From Systems Methodologies to an Age-Progressive Learning Architecture: The Conceptual Rationale for Systems-8

Integrating pedagogy, competency development, assessment, and research validation for young learners in the era of artificial intelligence

James Nguyen, Tinh Doan, and Ha Tran

Cybee Research Lab, Cybee Learning, Melbourne, Victoria, Australia

www.cybeelearning.com

Conceptual Working Paper, Version 3.08 (hypothesised specification)
| Draft for academic development and peer review

28 August 2026

Working-paper status and epistemic warning. Systems-8 is a *proposed* architecture, not validated yet. Every structural feature described in this paper – the eight domains, the four age bands, the six mastery levels, the twelve-component competency model – is a design hypothesis that the research programme in Section 15 is intended to test and, where necessary, refine, consolidate, or reposition. No claim of instructional effectiveness, developmental validity, or measurement quality is made or implied. Readers should treat Systems-8 as a research-stage specification whose structure and educational value remain subject to empirical validation. Public and commercial descriptions of the framework should remain aligned with the level of evidence available at each stage of the research programme.

Abstract

Artificial intelligence is changing the conditions under which education operates. Information, explanations, calculations, text, and code can increasingly be generated at low cost and high speed. This development does not make foundational knowledge less important; it makes the capacity to interrogate, connect, evaluate, and responsibly use knowledge more important. Systems thinking is relevant to this challenge because it directs attention to interdependence, purpose, boundaries, multiple perspectives, patterns over time, feedback, delay, unintended consequences, and adaptive intervention. Established systems methodologies and existing systems-thinking education and curriculum frameworks already provide substantial foundations. The gap advanced here is narrower: no located architecture simultaneously specifies systems-thinking competence as the organising construct, a multidimensional learner profile aggregating evidence across subjects, critical and normative systems concerns as assessed competencies, an assessment architecture for systems-thinking competence as such, and transfer to unfamiliar contexts as a separately assessed dimension of performance. This remains a falsifiable literature claim for Study 1.

Type of paper and approach. This is a conceptual paper of the *theory synthesis* type (Jaakkola, 2020): it integrates previously disconnected literatures – systems methodology, systems-thinking education, competency-framework design, educational measurement, and the transfer literature – into a proposed architecture, and specifies the empirical programme required to test it. It reports no new data and makes no effectiveness claim.

Proposition. Systems-8 is introduced as a pedagogical, competency, and assessment architecture rather than as a new systems methodology. It organises recurring systems concerns into eight inquiry domains: *Situation, Purpose, People and Perspectives, Parts, Wholes and Boundaries, Patterns, Connections and Emergence, Consequences,* and *Change*. These domains are intended to provide a common inquiry spine through which children can progressively access established systems ideas while the methodologies are preserved as theoretical sources, specialised toolsets, and advanced pathways. The paper distinguishes Systems-8 from existing systems-thinking education frameworks and national curriculum architectures (Section 6), states how the eight domains were derived and the rule under which they would be consolidated (Section 9), and confronts the transfer literature on which a central part of the architecture depends (Section 13). Twelve numbered propositions (P1–P12) are mapped to nine initial studies, with contingent follow-up studies triggered by pre-committed decision rules.

Conclusion. Systems-8 is a research proposition whose conceptual distinctiveness, developmental appropriateness, reliability, validity, transfer, cultural applicability, and incremental educational value must be tested empirically. The paper specifies how evidence should determine whether the architecture is retained, refined, integrated with stronger existing approaches, or repositioned.

Keywords: systems thinking; Systems-8; artificial intelligence; K-12 education; competency framework; educational assessment; transfer of learning; learning progressions; student agency

List of Abbreviations

Abbreviation	Full Form	Meaning / Relevance in This Paper
AI	Artificial Intelligence	Refers to computer systems capable of performing tasks associated with human intelligence. In this paper, AI is the broader technological context affecting learning, assessment, agency, and competence.
GPT-4	Generative Pre-trained Transformer 4	A large language model used in the Bastani et al. (2025) study discussed in relation to AI-assisted and unaided learning performance.
LLM	Large Language Model	A class of generative AI systems capable of producing and processing language-based content.
K–12	Kindergarten to Grade 12	Refers to school education from kindergarten through the final year of secondary schooling.
SD	System Dynamics	A systems modelling tradition focused on feedback, stocks, flows, delays, endogenous explanation, and behaviour over time.
SSM	Soft Systems Methodology	A systems methodology focused on problematic situations, purposeful activity, learning, and multiple worldviews.
CSH	Critical Systems Heuristics	A critical systems approach concerned with boundary judgements, legitimacy, power, expertise, inclusion, and affected stakeholders.
VSM	Viable System Model	A systems model concerned with organisational viability, autonomy, coordination, control, intelligence, and identity.
TOC	Theory of Constraints	A systems-oriented approach focused on identifying limiting constraints, causal logic, and focused improvement.
DSRP	Distinctions, Systems, Relationships, and Perspectives	A cognitive framework proposing four recurring patterns of thinking and used in this paper as an important comparator to Systems-8.

STH	Systems Thinking Hierarchical Model	A developmental systems-thinking framework proposed by Ben-Zvi Assaraf and Orion and discussed as one of the closest educational precedents to Systems-8.
SBF	Structure–Behavior–Function	A framework for analysing systems through their structures, behaviours, and functions.
NGSS	Next Generation Science Standards	US science education standards that include crosscutting concepts and grade-band progressions relevant to systems thinking.
NRC	National Research Council	The organisation responsible for major publications on K–12 science education and educational assessment cited in this paper.
ACARA	Australian Curriculum, Assessment and Reporting Authority	The authority responsible for the Australian Curriculum, used as a comparator for curriculum progression and systems-thinking content.
IB	International Baccalaureate	An international education organisation offering inquiry-based and transdisciplinary curricula.
PYP	Primary Years Programme	The International Baccalaureate programme for younger learners.
MYP	Middle Years Programme	The International Baccalaureate programme for middle-years learners.
OECD	Organisation for Economic Co-operation and Development	Source of the Learning Compass 2030 and related concepts of competence, knowledge, agency, anticipation, action, and reflection.
UNESCO	United Nations Educational, Scientific and Cultural Organization	Provides international guidance and competency frameworks relevant to education, sustainability, and artificial intelligence.
ILO	International Labour Organization	Provides labour-market evidence and estimates concerning occupational exposure to generative AI.
AERA	American Educational Research Association	One of the organisations responsible for the *Standards for Educational and Psychological Testing*.

APA	American Psychological Association	One of the organisations responsible for the *Standards for Educational and Psychological Testing*.
NCME	National Council on Measurement in Education	One of the organisations responsible for the *Standards for Educational and Psychological Testing*.
IRT	Item Response Theory	A psychometric approach used to analyse item difficulty, learner performance, and construct structure.
SOLO	Structure of the Observed Learning Outcome	A learning taxonomy used in this paper as a comparator for the proposed Systems-8 mastery progression.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses	A reporting standard for systematic reviews. The paper explicitly notes that the current conceptual review is not a PRISMA systematic review.
MIT	Massachusetts Institute of Technology	The university from which participants were drawn in the Booth Sweeney and Sterman stock-and-flow study discussed in the paper.
GCSE	General Certificate of Secondary Education	A UK secondary-school qualification referenced in Ofqual data on AI-related assessment malpractice.
AS	Advanced Subsidiary	A UK qualification level referenced together with GCSE and A-level data in the Ofqual evidence discussed in the paper.
OR	Operational Research	A field concerned with analytical and systemic approaches to decision-making and intervention; used in the context of community operational research.
P1–P12	Propositions 1–12	The twelve falsifiable propositions used to structure the proposed Systems-8 research programme.

Table of Contents

1. Introduction

Artificial intelligence is changing the relationship between knowledge, performance, and competence. Learners can now obtain explanations, summaries, calculations, code, images, and apparently polished arguments within seconds. Yet the availability of an answer does not establish that the learner understands the problem, can evaluate the evidence, recognises the answer's assumptions, or can anticipate the consequences of acting on it. In this paper, that condition is described as **performance without competence**.

This is not merely a conceptual worry. In a classroom-randomised field experiment with approximately 1,000 secondary students in Turkey, Bastani et al. (2025) found that access to an unmodified GPT-4 interface raised performance during assisted mathematics practice by 48%, but that the same students then scored 17% *worse* than controls on a subsequent unaided examination. A version with pedagogical guardrails avoided the same subsequent performance penalty, although it did not produce a learning gain relative to control. The largest current meta-analysis of experimental studies reports a substantial pooled positive effect of ChatGPT on measured academic performance (Hedges' $g = 0.712$, $k = 51$), while its authors caution that most primary studies used unproctored, immediate post-tests that may confound *output quality* with *learning gain*, and explicitly recommend proctored and delayed assessment (Deng et al., 2025). The divergence between these two bodies of evidence is precisely the measurement problem this paper takes as its starting point.

The appropriate response is not to reduce the importance of foundational learning. Literacy, numeracy, disciplinary knowledge, and self-regulation remain enabling conditions for more complex reasoning. The OECD Learning Compass explicitly combines knowledge, skills, attitudes, values, and student agency rather than treating generic competencies as substitutes for disciplinary knowledge (OECD, 2019a, 2019b, 2019c). UNESCO's guidance on generative AI and its AI competency framework for students likewise emphasise human agency, critical judgement, ethical responsibility, foundational AI knowledge, and responsible citizenship (Miao & Holmes, 2023; Miao et al., 2024). Children who lack substantive knowledge cannot reliably detect factual errors, evaluate causal claims, interpret data, or recognise when a fluent AI output is misleading. This paper's position is unambiguous on that point, and Section 13 returns to it in the strongest form the opposing literature can give it.

The challenge is therefore not whether children should continue learning mathematics, science, language, history, arts, and technology. They should. The challenge is whether education also develops the capacity to connect knowledge across domains, frame ambiguous situations, recognise interdependencies, examine who defines success, identify excluded perspectives, reason about patterns and feedback, anticipate consequences, and intervene responsibly under uncertainty. These capabilities are central to systems thinking (Meadows, 2008; Richmond, 1993; Sterman, 1994).

Systems thinking is not a single method or a universally agreed construct; rather, it encompasses multiple traditions with different theoretical and methodological commitments, including concerns with wholes, relationships, boundaries, purposes, structures, dynamics, multiple perspectives, and intervention (Midgley, 2000; Jackson, 2019, 2024). System dynamics focuses on endogenous causal structure, feedback, stocks, flows, delay, and behaviour over time (Forrester, 1994; Sterman, 2000). Soft systems methodology (SSM) addresses situations perceived as problematic in which people hold different worldviews, and in which learning about purposeful activity is more appropriate than presuming a single objective problem (Checkland & Poulter, 2020). Critical systems heuristics (CSH) examines boundary judgements, sources of control, expertise, legitimacy, and the position of those affected but not involved (Ulrich & Reynolds, 2020). The viable system model (VSM) addresses organisational viability, autonomy, coordination, control, intelligence, and identity (Beer, 1985). The theory of constraints (TOC) concentrates attention on limiting constraints, causal logic, and focused improvement (Goldratt, 1990). Critical systems thinking and systemic intervention emphasise methodological pluralism, context, power, ethics, and the consequences of boundary choices (Jackson, 2019, 2024; Midgley, 2000). Despite their different theoretical commitments and purposes, these traditions address overlapping systems concerns, including purpose, relationships, boundaries, structure, dynamics, perspectives, consequences, and intervention. Collectively, they provide the intellectual foundations relevant to this paper. As methodologies, however, they do not themselves specify how these concerns should be translated into an age-progressive pedagogy, competency progression, and assessment architecture for children.

Beyond the traditional methodologies, a second body of work already exists that is even more directly relevant: frameworks developed specifically to describe or teach systems thinking to school-age learners. These include the Systems Thinking Hierarchical model (Ben-Zvi Assaraf & Orion, 2005, 2010a), the Structure–Behavior–Function framework (Hmelo-Silver et al., 2007), Richmond's (1993) systems thinking skills, Arnold and Wade's (2015) systems thinking skill set, the Waters Center's fourteen Habits of a Systems Thinker (Waters Center for Systems Thinking, n.d.), and the systems-thinking competences embedded in sustainability frameworks (Bianchi et al., 2022; UNESCO, 2017; Wiek et al., 2011). Any proposal for a new architecture must be positioned explicitly against these, not only against the methodologies. Section 6 does so.

This creates the central theoretical problem the paper aims to answer, and it has two parts:

1. Why don't we simply create age-appropriate courses directly from the established methodologies?
2. Given that several systems-thinking education frameworks already exist – some with validated instruments – what does a new framework add?

The answer to both questions rests on a three-part gap. First, the established systems methodologies were developed mainly for professional, research, and intervention contexts rather than as developmental curricula for children. They do not specify what learners at successive ages should be able to demonstrate or how that development should be assessed. Second, the education frameworks reviewed in Section 6 each address important parts of the problem. These include cognitive progression, representation, systems skills, habits, curriculum concepts, and measurement. However, none provides the full educational architecture proposed here. Third, no located framework or curriculum architecture simultaneously specifies all five elements stated in Section 6.2. These are systems-thinking competence as the organising construct, a multidimensional learner profile that aggregates evidence across subjects, critical and normative systems concerns assessed as competencies, an assessment architecture for systems-thinking competence, and transfer to unfamiliar contexts assessed as a separate dimension of performance. The first two claims are supported by the review in Sections 5 and 6. The third remains a falsifiable literature claim that will be tested systematically in Study 1.

Systems-8 is therefore proposed not as another methodology competing with SSM, system dynamics, CSH, VSM, or TOC, nor as a replacement for existing education frameworks. It is proposed as a **pedagogical translation layer, a competency spine, and an assessment**

architecture positioned between established systems foundations and age-progressive educational practice.

These two questions define the central problem addressed by the paper. To answer them systematically, the current working paper is organised around five guiding questions. The first establishes the educational rationale, the second and third address the two central questions directly, the fourth develops the proposed architectural response, and the fifth specifies how that response should be tested. The five guiding questions are: (1) What educational and technological conditions make the explicit development of systems-thinking capabilities relevant for young learners? (2) Why do established systems methodologies, when taught independently, not by themselves, constitute an age-progressive educational architecture? (3) What do existing systems-thinking education frameworks and curriculum architectures already provide, and what remains unaddressed? (4) What kind of pedagogical, competency, and assessment architecture could integrate recurring systems concerns without replacing the established methodologies? (5) What programme of research is required to evaluate the conceptual validity, developmental appropriateness, assessment quality, transfer, and educational value of that architecture?

The intended contribution is fourfold. *Conceptually*, the paper distinguishes systems methodology, systems pedagogy, systems competence, and systems assessment, and locates existing frameworks within that distinction. *Architecturally*, it proposes an integrated structure linking inquiry, age progression, observable outcomes, and learner evidence. *Critically*, it states the framework's load-bearing assumption about transfer and confronts the literature that disputes it. *Methodologically*, it specifies twelve falsifiable propositions and the studies that would test them.

Structure of current working paper. Section 2 states the method by which the conceptual argument was constructed. Sections 3 and 4 establish the educational and technological case. Section 5 reviews the systems-thinking methodologies; Section 6 reviews the existing education frameworks and states the differentiation claim. Sections 7 and 8 develop the argument for an additional architectural layer. Sections 9 to 12 specify the framework. Section 13 engages directly with the transfer critique. Section 14 states traceability to the methodological foundations. Section 15 sets out the propositions and the research programme; Section 16 the challenges and safeguards. Sections 17 and 18 discuss and conclude.

2. Method: how this conceptual argument was constructed

Conceptual papers are theory-building research and require an explicit and defensible design rather than an unstructured essay form (Jaakkola, 2020; MacInnis, 2011). This section states the design so that readers can evaluate the argument's coverage and identify what it may have missed.

Type of conceptual contribution. In Jaakkola's (2020) typology, this paper is primarily a *theory synthesis*: it integrates literatures that are rarely brought together – systems methodology, school-age systems-thinking education, competency-framework design, educational measurement theory, and the transfer literature – to produce an architecture none of them individually supplies. In MacInnis's (2011) terms the contribution is *integrating* and *delineating* rather than *identifying* a new construct. This matters for how the paper should be judged: the test of a synthesis is whether the integration is defensible and generative, not whether its component concepts are novel. Systems-8 does not claim that situation, purpose, stakeholders, boundaries, patterns, feedback, consequences, or intervention are new ideas.

Literature selection. This conceptual paper draws on four literature streams. The selection was purposive and iterative rather than systematic. Study 1, described in Section 15, is intended to provide the systematic review needed to test and revise these boundaries.

(a) Systems methodology. Widely cited foundational and synthesising sources were selected from the major systems traditions most relevant to the conceptual problem, together with key treatments of methodological pluralism. The selection drew partly on prior knowledge of the field rather than on a comprehensive database search. Which traditions and sources are treated as foundational is itself a boundary judgement, so this selection should be regarded as provisional rather than exhaustive. Study 1 will revisit that boundary systematically.

(b) Systems thinking in school-age education. Targeted searches and citation tracking were used to identify sources that proposed frameworks of systems-thinking components, reported empirical evidence of school-age learners' systems reasoning, or presented validated measurement instruments. The purpose was to identify the strongest relevant educational precedents rather than to claim exhaustive coverage.

(c) Competency and assessment. The paper drew on established work in educational measurement, validity theory, competency modelling, and learning-progression methodology. These sources were selected for their relevance to the design and evaluation of the proposed competency and assessment architecture.

(d) Counter-evidence and competing alternatives. Literature that could challenge the proposal was sought deliberately. This included work questioning the teachability and transfer of domain-general thinking skills, as well as existing frameworks or curriculum architectures that might already provide some or all of what Systems-8 proposes. These two lines of challenge correspond to the objections addressed later in the paper: whether the proposed competencies can transfer beyond specific disciplinary contexts, and whether a new architecture is needed at all.

Boundary of the review. The review is *purposive and integrative*, not systematic in the PRISMA sense. It does not claim exhaustive coverage. Three consequences follow and should be held against the paper. Frameworks published in languages other than English are under-represented, and this is a known rather than a hypothetical gap: the review did not cover Vietnamese-language sources, despite Vietnam being an initial target context. Practitioner frameworks circulated as grey literature are captured unevenly. And the search was conducted by parties who have an interest in the proposal succeeding. Study 1 of the research programme (Section 15) is designed as the systematic review this paper does not perform, and is explicitly empowered to overturn the framework's structure. Its scope is specified to include four categories the present review did not systematically cover: cognitive frameworks of systems thinking outside the education literature; national and international curriculum architectures that embed systems concepts as a continuous progression; Vietnamese-language sources and programme records; and measurement instruments rather than frameworks alone. The omission that motivated this specification was categorical rather than incidental, and it is documented in Section 6.2.

Argumentative strategy. The paper adopts a strongest-objection-first structure. The two objections most likely to defeat the proposal are: (1) that it is a rebranding of existing frameworks (Section 6), and (2) that its transfer assumption is contradicted by the evidence (Section 13).

What this paper does not do. It reports no empirical data. It makes no claim about instructional effectiveness. It does not report psychometric properties, because none exist. It does not claim that the eight domains are the correct or only decomposition of systems inquiry.

3. The educational context: why systems thinking matters

The case for systems thinking begins with the nature of the situations learners are increasingly likely to encounter. Many contemporary social, technological, organisational, and environmental problems cannot be adequately understood by treating variables as isolated or relationships as stable. Their behaviour may arise from interactions among multiple actors, feedback, delays, adaptation, path dependence, incomplete information, and effects that unfold across time and social groups (Sterman, 1994; Midgley, 2000; Meadows, 2008). Climate adaptation, public health, platform governance, migration, urban development, biodiversity, misinformation, and AI-enabled decision systems all exhibit some of these characteristics. Education focused primarily on pre-defined and well-structured problems may therefore leave learners less prepared for situations in which stakeholders disagree about what the problem is, what counts as improvement, and which concerns should be included (Rittel & Webber, 1973; Checkland & Poulter, 2020; Ulrich & Reynolds, 2020).

One reason this matters is that people often find dynamic behaviour difficult to reason about. Sterman (1994) argued that people tend to focus on events, underestimate feedback and delays, and apply open-loop reasoning to situations whose behaviour is generated by closed feedback processes. Booth Sweeney and Sterman (2000) provided empirical evidence of this difficulty using short stock-and-flow tasks with MIT graduate students. These tasks assess whether participants understand accumulation over time. In the well-known bathtub task, participants are shown changing rates of water flowing into and out of a bathtub and are asked to infer how the amount of water in the tub changes. The underlying principle is simple: the amount of water increases when inflow exceeds outflow and decreases when outflow exceeds inflow. Despite the apparent simplicity of the task, average performance was about 77% on the basic version and fell to roughly 48% on a more difficult variant. Performance did not vary systematically with prior education in mathematics or science (Booth Sweeney & Sterman, 2000). The educational implication should be stated cautiously. These results do not show that general quantitative education is ineffective, but they do suggest that stock-and-flow reasoning is not necessarily acquired through quantitative education alone. Section 13.3 returns to this finding because it constrains claims about transfer as much as it supports the case for teaching systems reasoning explicitly.

The challenge extends beyond accumulation. Systems thinking directs attention away from isolated events towards patterns of behaviour and the structures that help generate them (Richmond, 1993; Meadows, 2008). It also shifts attention from individual variables towards interrelationships, feedback, and change over time (Sterman, 1994, 2000), and from immediate outcomes towards delayed and wider consequences within a system (Meadows, 2008; Sterman, 2000). These forms of reasoning are important because interventions that appear sensible in the short term can generate counterintuitive or unintended outcomes when feedback, delays, and leverage points are misunderstood (Meadows, 2008).

In social systems, however, the challenge is not only technical. Critical systems traditions emphasise that analysts and participants make boundary judgements about what is relevant, who should be included, what counts as improvement, and whose knowledge is treated as legitimate

(Midgley, 2000; Ulrich & Reynolds, 2020). A solution may appear coherent within one boundary while producing adverse consequences that become visible only when the boundary is widened. Systems reasoning therefore involves more than understanding causal structure. It also requires attention to purpose, participation, exclusion, values, and the distribution of consequences.

This perspective also changes how models should be treated in education. Systems models are not complete descriptions of reality. They are purposeful representations shaped by choices about relevance, evidence, assumptions, and system boundaries (Midgley, 2000; Ulrich & Reynolds, 2020). Producing a sophisticated diagram should therefore not, by itself, be treated as evidence of systems understanding. A causal-loop diagram, stakeholder map, behaviour-over-time graph, or other system representation becomes educationally meaningful when the learner can explain why particular elements were included, justify the relationships shown, identify important exclusions and assumptions, and revise the representation when new evidence or perspectives emerge. This combination of representation and justification is particularly important for Systems-8 because the proposed assessment architecture is intended to evaluate reasoning rather than simply the production of system artefacts.

There is evidence that selected elements of systems thinking can be developed before university, although the evidence remains fragmented and varies considerably in scope and strength. Fisher (2023) provides a broad background, documenting the use of systems-thinking activities across K-12 settings over two decades, including stories, behaviour-over-time graphs, causal diagrams, and system dynamics modelling. Evidence involving younger learners includes Feriver et al. (2019), who found that children aged four to six could demonstrate emerging systems-thinking abilities when tasks were concrete and developmentally appropriate, although capability varied considerably. At elementary-school level, Hokayem et al. (2020) showed that learners could reason about feedback loops, while also finding that such reasoning was difficult and depended on knowledge sources and instructional context. Among fourth-grade students aged approximately nine to ten, Ben-Zvi Assaraf and Orion (2010a) found that an inquiry-based earth-systems curriculum supported movement from limited decomposition of a system towards identifying components, processes, and interconnections, although only some students reached the higher levels of their proposed hierarchy. With older primary-school learners, Mambrey et al. (2020) used a 36-item instrument analysed with item response theory with 196 German fifth- and sixth-grade students. They found that distinguishing direct from indirect causal effects was a major source of difficulty and that, once indirect relationships were understood, additional system complexity added comparatively less difficulty. At seventh-grade level, Raved and Yarden (2014) found that systems-oriented instruction supported students in developing a more integrated understanding of the human circulatory system. These studies involve learners of different ages and examine systems reasoning in different disciplinary contexts, using different instructional approaches and measures. Their ordering here by learner age is descriptive only and does not constitute evidence of a common developmental sequence.

Taken together, these studies support a narrower conclusion than is sometimes drawn from the systems-education literature. They indicate that school-age learners can engage with selected systems concepts when instruction, representation, and content are appropriately designed. They do not establish a universal developmental sequence for systems thinking. They also do not demonstrate that competence developed in one subject automatically transfers to another. These limitations are important for Systems-8. The framework therefore treats age progression and transfer as propositions to be tested rather than as established facts. At the same time, the existing evidence provides sufficient reason to reject the assumption that systems reasoning must be deferred until adulthood.

The relevance of systems thinking is also broadly consistent with contemporary educational policy. The OECD Learning Compass 2030 frames education around individual and collective well-being, transformative competencies, and a cycle of anticipation, action, and reflection (OECD, 2019a). Student agency is described as the capacity and willingness to influence one's life and surroundings responsibly (OECD, 2019c). Systems thinking may support this orientation because anticipation involves considering relationships and possible consequences, action involves intervention in interconnected situations, and reflection involves reconsidering assumptions in light of outcomes. The OECD framework provides a relevant policy context for these capabilities. It should not be treated as evidence for Systems-8 or for any particular systems-thinking architecture.

Systems thinking should also not be treated as sufficient on its own or as a substitute for disciplinary knowledge and other educational capabilities. A learner may construct a technically coherent causal model while accepting an unjust purpose, excluding affected groups, overlooking weak evidence, or failing to revise the model when circumstances change. For this reason, the proposed Systems-8 architecture combines its eight core systems competencies with the four cross-cutting capabilities specified in Section 11.2: Evidence and Epistemic Discipline, Metacognition and Model Revision, Communication and Collaboration, and Ethical Responsibility and Agency. These capabilities are intended to shape the quality with which systems reasoning is enacted rather than to replace subject knowledge. The educational case for Systems-8 therefore rests on the more limited proposition that explicit systems capabilities, combined with disciplinary knowledge and these cross-cutting capabilities, may help learners reason more effectively about interconnected, dynamic, and contested situations.

4. AI-enabled conditions and the risk of performance without competence

AI can now reproduce many conventional educational outputs. When it does, several inherited patterns in education become more consequential. The claim is not that current education should be replaced. Existing education systems deliver indispensable benefits, including literacy, numeracy, disciplinary knowledge, socialisation, and civic preparation. The risks below are structural tendencies rather than universal descriptions of every school or teacher. Six risks are considered below. Several concerns that are often listed separately are grouped under Risk 1 because they reflect the same underlying construct-validity problem.

Risk 1: When Assessment No Longer Measures the Intended Construct. This is the central assessment risk in AI-rich education. It includes several concerns that are often treated separately, such as overvaluing answer production, mistaking fluent output for understanding, and relying on unsupervised written work as evidence of individual competence. If assessment focuses mainly on the final product, the observed performance may reflect the combined capability of learner and tool rather than the intended competence of the learner. The problem is therefore not primarily academic misconduct. It is a problem of construct validity in the technical sense (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Kane, 2013; Messick, 1995).

Evidence of widespread generative-AI use makes this concern increasingly difficult to ignore. Chirikov et al. (2026), surveying more than 95,000 students at twenty US public research universities and using list randomisation to reduce social-desirability bias, found that about 37% used generative AI monthly or more and that 9% reported using it to cheat, rising to 26% among daily users. By contrast, Ofqual (2025) recorded 100 detected cases of AI-related plagiarism nationally across approximately 1.38 million GCSE, AS and A-level candidates in summer 2025. These figures come from different populations, age groups, assessment settings, definitions, and detection systems, so they should not be interpreted as comparable rates. Their divergence nevertheless illustrates a broader problem. Under conditions of widespread AI use, the inference from a polished product to the competence of an individual learner becomes harder to justify. Detection tools cannot resolve this problem because an AI-assisted product may be original in wording while still concealing how much of the reasoning was performed by the learner.

For systems-thinking education, the implication is not that AI must be excluded from assessment, but product quality alone is insufficient evidence of systems competence. Learners should also be able to explain the relationships they have represented, justify their boundary choices and assumptions, identify relevant and excluded perspectives, anticipate consequences, revise a model when challenged, and demonstrate related reasoning in a different context. These are not merely assessment safeguards. They are part of the systems capabilities that the curriculum is intended to develop. Systems-8 therefore distinguishes supported performance from unaided competence and treats reasoning, justification, revision, and transfer as evidence alongside the final artefact.

Risk 2: AI-supported cognitive offloading may substitute for learning-relevant effort. External tools have always extended human cognition. Writing, calculators, maps, and search engines can reduce memory or processing demands, and such cognitive offloading is not inherently harmful (Risko & Gilbert, 2016). The educational issue is therefore not whether learners offload cognitive work, but what is offloaded and whether the tool complements or substitutes for the activity through which learning occurs. Generative AI extends this issue because a language model can support several stages of a task at once, including interpreting a question, proposing a solution, generating an explanation, and revising the final response. In some circumstances this may free attention for higher-order reasoning. In others, it may reduce engagement with the reasoning that the learner is expected to develop. The broader learning literature similarly shows that conditions requiring productive effort can support later learning, although difficulty is beneficial only when it is appropriately designed and manageable for the learner (Bjork & Bjork, 2011).

Direct evidence in school-age learners is beginning to show why this distinction matters. Bastani et al. (2025) conducted a classroom-randomised experiment with approximately 1,000 secondary-school mathematics students in Turkey. Students given unrestricted access to GPT-4 performed substantially better during AI-assisted practice, but subsequently performed worse than the control group when completing an unaided examination. A version of the AI tool designed with pedagogical guardrails preserved much of the practice benefit without the same subsequent performance penalty. The study therefore does not show that AI use generally harms learning. It shows that *how AI assistance is designed can determine whether improved supported performance is accompanied by comparable unaided learning.*

Evidence from older learners points in the same direction but is more mixed. Lehmann et al. (2025), using two pre-registered experiments and a field study with university students, found no significant overall learning disadvantage from access to large language models. Their exploratory analyses nevertheless suggested that outcomes differed according to how the tool was used. Learners who used the model mainly to generate solutions could cover more material but showed weaker understanding of individual topics than learners who used it to request explanations and support their own reasoning. Gerlich (2025) reported associations among greater AI use, cognitive offloading, and lower critical-thinking performance in participants aged seventeen and above. Because the study was cross-sectional, it cannot establish that AI use caused the observed differences or that the same relationship applies to children. Preliminary work such as Kosmyna et al. (2025) may add further evidence about changes in cognitive engagement during LLM-assisted tasks, but it should not yet be treated as evidence of developmental effects.

The evidence therefore supports a narrower claim than a stronger developmental interpretation would warrant. AI assistance can alter the amount and type of cognitive work performed by the learner, and supported performance may not correspond to what the learner can subsequently do unaided. Direct evidence that extensive AI substitution during childhood prevents a capability from forming does not yet exist. That possibility should therefore be treated as a research question rather than an established effect. For Systems-8, the implication is to distinguish between AI that *complements* systems reasoning and AI that *substitutes* for it. Learning activities should require learners to construct, explain, justify, challenge, and revise their own reasoning, while assessment should make clear what can be demonstrated with AI support and what can be demonstrated independently. UNESCO's emphasis on age-appropriate, human-centred use of generative AI is consistent with this precautionary approach, but it should be treated as policy guidance rather than empirical evidence of developmental harm.

Risk 3: Knowledge may remain fragmented across subjects. Disciplinary organisation is necessary because fields possess distinctive concepts, evidence standards, and methods. The risk arises when specialisation is not accompanied by integration. Real situations do not arrive divided into school subjects: an AI-supported health policy decision may require biological knowledge, statistics, ethics, law, economics, public administration, and community perspectives. Generative AI sharpens this risk, because it supplies disciplinary facts on demand while leaving the work of connecting them to the learner. The OECD distinction among disciplinary, interdisciplinary, epistemic, and procedural knowledge recognises this need (OECD, 2019b). Systems thinking offers a structure for connecting these forms of knowledge without dissolving their disciplinary integrity.

Risk 4: Learners may receive insufficient practice in problem framing and may encounter repeated linear causal structures. Many conventional classroom tasks present the problem, variables, goals, and evaluation criteria in advance. In professional and civic life, competing actors may disagree about the problem, its causes, the relevant boundary, and what constitutes success. Checkland and Poulter (2020) emphasise that human activity situations cannot always be reduced to an agreed problem awaiting optimisation. A learner who cannot interrogate the initial frame may solve the wrong problem efficiently, and AI can sharpen this risk when it answers the question as asked without challenging how it was framed. Classroom simplification has a second effect. One cause, one effect, one method, and one answer can be useful simplifications for teaching basic concepts, but repeated reliance on them may encourage the assumption that social and organisational problems behave similarly. Feedback means that effects later alter causes; delays separate action from visible result; people adapt to interventions; and local improvement can degrade system performance elsewhere (Meadows, 2008; Sterman, 2000). In Mambrey et al. (2020), distinguishing direct from indirect causal effects was a major source of difficulty for the fifth- and sixth-grade learners in their sample.

Risk 5: Student agency may be weakened. A strongly instruction-centred learning environment may provide fewer opportunities for learners to exercise agency by defining purposes, making choices, challenging assumptions, considering alternative perspectives, and acting under uncertainty. This concern is consistent with the OECD Learning Compass, which treats student agency as the capacity and willingness to influence one's life and surroundings responsibly (OECD, 2019c). The issue becomes more salient when algorithmic systems increasingly mediate information, recommendations, attention, and choice. Generative AI extends this mediation because it can propose not only an answer, but also the framing of the problem, relevant assumptions, possible perspectives, and courses of action.

Technological proficiency alone therefore does not ensure that learners can judge purposes, question assumptions, consider who may be affected, or decide when a technological recommendation should be challenged. UNESCO's AI competency framework for students similarly treats AI competence as broader than tool use. It includes four competency aspects: a human-centred mindset, ethics of AI, AI techniques and applications, and AI system design, organised across three progression levels (Miao et al., 2024).

For Systems-8, the implication is that student agency should be practised rather than merely described. Learners should have opportunities to frame problems, question given purposes, compare perspectives, make and justify boundary choices, evaluate proposed interventions, and revise their position when new evidence or consequences emerge. This rationale is reflected particularly in the cross-cutting capability of *Ethical Responsibility and Agency* described in Section 11.2. The aim is not to reduce learners' use of AI, but to ensure that AI support does not replace their responsibility for judgement and action.

Risk 6: Educational systems may optimise measurable outputs while losing sight of purpose.

Marks, completion, attendance, and rankings are necessary administrative signals, but when measures become dominant targets, participants may adapt behaviour to improve the metric rather than the underlying purpose, a problem long recognised in the social-indicator literature (Campbell, 1979). Generative AI can intensify this risk by making some visible outputs easier to produce without equivalent evidence of the learning those outputs were intended to represent. Where Risk 1 concerns the validity of an individual score, this risk concerns the incentives created by the system that aggregates and rewards such scores. A changing labour market adds context rather than proof. The International Labour Organization's refined global index indicates that roughly one in four workers is in an occupation with some generative-AI exposure, although only about 3.3% of global employment falls in the highest-exposure band; the authors emphasise transformation rather than replacement (Gmyrek et al., 2025). A subsequent ILO synthesis cautions that competing exposure measures correlate with wages in directions ranging from −0.56 to +0.54 and cannot be interpreted as predictions of job displacement (Merola et al., 2026). The World Economic Forum (2025) reports employer expectations of substantial skill change by 2030, with analytical thinking, resilience, flexibility, leadership, creative thinking, and technology-related capabilities remaining important. These are exposure indices and employer expectations, not outcome data, and they do not establish which curriculum is needed. They do, however, reinforce the relevance of adaptability and judgement under changing conditions.

Taken together, these six risks do not by themselves establish a case for Systems-8. They identify a set of educational design requirements that become more important under AI-enabled conditions. Assessment should distinguish supported performance from what learners can demonstrate independently. AI should complement rather than substitute for the reasoning involved in learning. Disciplinary knowledge should remain strong while learners are also given opportunities to connect knowledge across subject boundaries. Learners need practice in framing problems, reasoning about relationships, feedback, and consequences, and revising their understanding when conditions change. They also need opportunities to exercise agency and ethical judgement rather than simply follow a given problem definition or technological recommendation. Finally, educational measures should remain connected to their underlying purposes rather than becoming ends in themselves.

These requirements point towards an integrative educational capacity that complements rather than replaces disciplinary teaching, teachers, and structured practice. They do not, however, demonstrate that a new framework is necessary. The next question is whether established systems traditions and existing systems-thinking education frameworks already provide the intellectual and educational resources required. Sections 5 and 6 examine that question before any additional architecture is proposed.

5. Established systems foundations and their educational value

Systems-8 begins from the recognition that systems thinking already has extensive theoretical and methodological traditions. Its rationale cannot rest on concepts that the field has already developed, but on how those concepts are translated into a coherent developmental educational architecture.

System dynamics (SD): offers a powerful account of dynamic complexity. Its emphasis on behaviour over time, endogenous explanation, feedback, stocks, flows, and delays helps learners move beyond event-based explanations (Forrester, 1994; Sterman, 2000). In education, behaviour-over-time graphs and causal-loop diagrams can make reasoning visible and allow learners to test explanations. Yet system dynamics also requires care: a child may learn the notation of reinforcing and balancing loops without understanding evidence, boundary choice, or the social legitimacy of the model. Richmond (1993), writing from within this tradition, named seven distinguishable thinking skills – dynamic, system-as-cause/closed-loop, generic, structural, operational, continuum, and scientific thinking – and in doing so produced what is arguably the first explicit systems-thinking *skills* framework. Group model building extended the tradition towards participatory learning with stakeholders (Vennix, 1996), a precedent directly relevant to collaborative classroom work.

Soft systems methodology (SSM): contributes a different orientation. Rather than assuming an objectively given problem, SSM explores situations perceived as problematic, uses rich pictures and purposeful activity models, and treats inquiry as a learning process among actors with different worldviews (Checkland & Poulter, 2020). Its educational value lies in resisting premature problem definition and recognising that different actors may hold legitimate but incompatible interpretations. However, SSM was developed through professional action research and is not itself an age-progression curriculum.

Critical systems heuristics (CSH): extends the analysis of boundaries and legitimacy. It asks who ought to be considered client, decision maker, expert, witness, and affected party; what sources of motivation, control, knowledge, and legitimacy shape the intervention; and how claims change when boundaries change (Ulrich & Reynolds, 2020). This is particularly important for AI education, because technical systems routinely embody decisions about whose data, needs, and values count. Yet CSH is conceptually and normatively demanding and requires careful pedagogical translation for younger learners.

The viable system model (VSM): examines how organisations remain viable through distributed operations, coordination, control, intelligence, and policy or identity (Beer, 1985), supporting reasoning about autonomy, governance, and recursive organisation. *The theory of constraints* directs attention to the limiting factor that constrains system performance and promotes focused improvement through causal logic (Goldratt, 1990). *Complexity perspectives* contribute emergence, adaptation, nonlinearity, and sensitivity to context. *Critical systems thinking and systemic intervention* argue that no method is universally appropriate and that methodology selection is itself an ethical and political act (Jackson, 2019, 2024; Midgley, 2000).

These approaches should not be flattened into interchangeable tools. They differ in ontology, epistemology, purpose, and intervention logic. System dynamics commonly seeks endogenous explanations of behaviour; SSM treats worldviews and purposeful activity as central; CSH foregrounds boundary judgement and legitimacy; VSM focuses on viability; TOC focuses on constraints and leverage. Their differences are educational assets, because advanced learners should eventually understand methodological choice and pluralism. The purpose of Systems-8 is not to erase these differences but to create a common developmental route through which learners can progressively encounter them.

6. Existing systems-thinking education frameworks: what already occupies this space

Existing systems-thinking education frameworks and curriculum architectures provide substantial foundations for the present proposal. They offer developmental hierarchies, representations of systems reasoning, systems-thinking skills and habits, curriculum progressions, and, in some cases, validated instruments or formal assessment architectures. Systems-8 is intended to build on this work rather than replace it. Its proposed contribution lies in bringing together five elements within one educational architecture: systems-thinking competence as the organising construct; a multidimensional learner profile that aggregates evidence across subjects; critical and normative systems concerns operationalised as assessed competencies; a specified assessment architecture for systems-thinking competence; and transfer to unfamiliar contexts as a separately assessed dimension of performance. Each of these elements has precedent in the existing literature, but no framework or curriculum architecture located in the present review combines all five. This remains a provisional differentiation claim to be tested systematically in Study 1.

6.1 Existing frameworks

The Systems Thinking Hierarchical (STH) model (Ben-Zvi Assaraf & Orion, 2005) is the closest prior art and must be addressed directly, including an uncomfortable coincidence: it comprises **eight** emergent characteristics, arranged in three sequential levels – analysis of system components; synthesis of components (identifying relationships, dynamic relationships, organising within a framework of relationships, understanding cyclicity); and implementation (recognising hidden dimensions, thinking temporally through retrospection and prediction, generalising). It was developed with eighth-graders in earth-systems education and extended downward to fourth-graders (Ben-Zvi Assaraf & Orion, 2010a) and forward six years longitudinally (Ben-Zvi Assaraf & Orion, 2010b). It is genuinely developmental, genuinely school-based, genuinely evidenced, and genuinely eight-part. Any claim that no age-progressive systems-thinking model exists is false, and this paper does not make it.

What the STH model is not is a *curriculum-wide competency and assessment architecture*. It is a hierarchy of cognitive characteristics elicited through repertory grids, concept maps, drawings, and interviews within earth-systems content; it does not specify cross-subject learning outcomes, does not include boundary judgement, purpose contestation, stakeholder legitimacy, or ethical agency, and its evidence base sits within a single subject area in a single national context.

DSRP (Cabrera et al., 2008; Cabrera & Colosi, 2008) is the framework closest to Systems-8 in component vocabulary. It was developed primarily as a cognitive framework rather than as a school curriculum and assessment architecture, though it has been taken up in schools. It identifies four patterns – distinctions, systems, relationships, and perspectives – held to underlie cognition generally, with systems thinking treated as an emergent property of their deliberate and recursive application. The resemblance to four of the present framework's eight domains is close, and readers will make the comparison whether or not this paper invites it.

The two are objects of different kinds. DSRP is a theory of the patterns underlying thought. Systems-8 is a developmental educational architecture that uses inquiry domains to organise curriculum, competence, tasks, and assessment. The comparison should therefore be made at the level of what each specifies rather than at the level of vocabulary. Systems-8 is not proposed as a more correct set of fundamental systems concepts than DSRP; it serves a different architectural purpose.

The introduction of DSRP was accompanied by a substantive disagreement about its relationship to methodological pluralism. Setting itself explicitly against the pluralism it attributes to Gregory, Jackson, Midgley, and White and Taket, the founding article identifies "four universal conceptual patterns that apply to all human thinking and thus crosscut systems models and systems thinking" (Cabrera et al., 2008, p. 299). Midgley (2008) responded directly, arguing that DSRP's attempt to rationalise and unify systems thinking should not displace methodological pluralism, and that DSRP itself shares important theoretical ground with pluralist traditions while lacking the plurality of methods available within them. The disagreement is noted here because the present paper's position depends on neither side of it. Systems-8 makes no claim of the kind contested in that exchange: the eight domains are not offered as a universal ontology of thought, nor as the correct decomposition of systems inquiry, but as a provisional educational decomposition stated together with the derivation that produced it (Section 9.2), the collapse rule that would reduce it, and the studies that could overturn it. This is deliberately narrower than the universalising claim debated in that exchange, and that narrower scope is what makes the framework testable rather than definitional.

Claims are made for educational outcomes associated with DSRP-based instruction in schools, including effects on problem solving, pro-social behaviour, and attainment in other subjects. A peer-reviewed study reporting such outcomes with a comparison condition, specified instruments, and delayed assessment could not be located during the present review. The founding statement of the theory is Cabrera's (2006) Cornell doctoral dissertation, which presents it as a conceptual framework offered as a first step and states explicitly that future research is required to test it. That dissertation is also worth citing for a second reason, unrelated to the comparison: its methodological review of the systems-thinking literature found a disproportionate number of descriptive rather than empirical studies, and construct-validity problems in the empirical designs that did exist. That diagnosis, reached independently and from within the closest comparator tradition, corroborates the assessment gap stated in Section 8 rather than contradicting it. Should evidence of school outcomes exist and hold, it would bear directly on P8 and P9 and the positioning taken here would require revision; Study 1 is instructed to resolve this.

The Structure–Behavior–Function (SBF) framework (Hmelo-Silver et al., 2007) distinguishes the visible parts of a system, the causal mechanisms by which parts achieve outcomes, and the purposes of the system and its parts. The expert–novice gap lies almost entirely in behaviours and functions rather than structures – a finding with direct pedagogical implications. SBF is a powerful analytic lens with a substantial research base, and it is deliberately narrow: it is a representation scheme for understanding a system, not a framework for framing contested situations, weighing stakeholder legitimacy, or designing intervention.

Richmond's (1993) seven systems thinking skills constitute the first explicit skills framework in the field and remain influential within system dynamics education. They are, however, formulated for and by system dynamicists, are strongly oriented to stock-flow and feedback reasoning, and carry no age progression and no assessment model.

Arnold and Wade's (2015, 2017) systems thinking skill set defines systems thinking through a systems approach and specifies eight skills including recognising interconnections, understanding feedback, differentiating stocks and flows, identifying non-linear relationships, and understanding systems at different scales. It is rigorous and widely cited, but it was developed for systems engineering practice, targets adult professionals, and offers no student assessment instrument.

The Waters Center's fourteen Habits of a Systems Thinker (Waters Center for Systems Thinking, n.d.; Benson & Marlin, 2020) is one of the closest practitioner comparators to Systems-8 in pedagogical orientation. Its habits include surfacing and testing assumptions, changing perspectives, considering short- and long-term consequences, recognising interdependencies, and checking results in order to adjust actions. These concerns overlap substantially with several Systems-8 domains and cross-cutting capabilities. The similarity is therefore substantive and should be acknowledged rather than minimised.

The difference lies less in the concepts themselves than in what the two frameworks are designed to specify. The Habits provide a practical repertoire for cultivating systems-oriented ways of thinking and are supported by age-differentiated classroom materials. In the published and practitioner materials located for the present review, however, they are not organised as a multidimensional learner competency profile with empirically tested developmental levels, a construct-specific assessment architecture, or transfer to unfamiliar contexts as a separately assessed dimension of performance. No peer-reviewed validation study or published psychometric evidence for a Habits-based measurement instrument was located. Systems-8 therefore does not claim to replace or improve the Habits as a practitioner framework. Its proposed contribution is to build an explicit competency, progression, evidence, and assessment architecture around overlapping systems concerns, a claim that remains to be tested empirically.

Sustainability competence frameworks name systems thinking as a headline competence. Wiek et al. (2011) place systems-thinking competence first among five key sustainability competencies for higher education. UNESCO (2017) lists systems thinking competency first among eight cross-cutting competencies for education for sustainable development. GreenComp (Bianchi et al., 2022) places systems thinking as competence 2.1 within its "embracing complexity" area, across four areas and twelve competences. These are authoritative and policy-relevant. They are also, by design, *normative reference frameworks*: they define what should be developed without specifying age progression, pedagogy, task design, or measurement, and none carries a validated instrument.

National and international curriculum architectures are essential comparators, because they already provide continuous curriculum progressions and, in some cases, associated assessment specifications. The *Next Generation Science Standards* specify seven crosscutting concepts – patterns; cause and effect; scale, proportion and quantity; systems and system models; energy and matter; structure and function; and stability and change – together with grade-band progressions describing how each develops across K-12 (National Research Council, 2012;

NGSS Lead States, 2013). The standards are three-dimensional: performance expectations combine disciplinary core ideas, science and engineering practices, and crosscutting concepts, and *Developing Assessments for the Next Generation Science Standards* (National Research Council, 2014) specifies a corresponding assessment architecture. This is a continuous, official, assessed progression, and it is a stronger prior architecture than any of the frameworks reviewed above.

Its scope, however, is science and engineering. The crosscutting concepts are transdisciplinary within those fields rather than across the whole curriculum, and they are not aggregated into a systems-thinking competency profile that accumulates evidence from history, literature, civics, or everyday social contexts. Nor do the standards develop the critical and normative concerns – who is entitled to define purpose, whose knowledge counts as legitimate, who is affected but not involved – that critical systems traditions place at the centre of inquiry.

The *Australian Curriculum* specifies systems thinking within Design and Technologies from Foundation to Year 10, with content descriptions and achievement standards at each band, and states that students apply design and systems thinking, evaluate and iterate solutions, consider social and environmental impacts, make informed and ethical decisions, and transfer knowledge to new situations. Its Sustainability cross-curriculum priority is organised around three concepts – systems, world views, and futures – and addresses interdependence among social, economic, and ecological systems, diversity of values, equity and social justice, and consequences across generations. It is substantially stronger on critical and normative concerns than a review of systems-thinking education frameworks alone would suggest. What it does not supply is a systems-thinking construct model that is progressed and reported independently of learning areas, a common learner profile accumulating evidence across subjects, or a specification of how such competence would be scored.

The *International Baccalaureate* Primary Years and Middle Years Programmes provide inquiry-based, transdisciplinary, concept-driven curricula for learners from early childhood, and are a significant comparator for the proposition that conceptual learning can be organised across subjects and across ages. They do not organise learner progression around a defined systems-thinking construct, and no validated systems-thinking profile associated with them could be located.

Validated school-age instruments are the rarest category. Mambrey et al. (2020) developed a 36-item multiple-choice instrument analysed with item response theory on 196 fifth- and sixth-graders, supporting a three-dimensional model – identifying system organisation, analysing system behaviour, and system modelling. This is among the strongest psychometric evidence in the school-age systems-thinking literature, and it is worth noting what it found: *three* dimensions, not eight. Verhoeff et al. (2018) argue on theoretical grounds that skill-list framings of systems thinking, including the STH model, are behaviourally shallow and should be grounded in systems-theoretical characteristics.

The critique by Verhoeff et al. (2018) applies directly to Systems-8 and cannot be answered merely by expanding a list of skills. The proposed architecture therefore does not define competence as the presence of isolated behaviours or the correct use of named tools. Each core competency is anchored to a systems-theoretical concern; evidence must include justification of relationships, assumptions, and boundaries rather than performance of a technique alone; and the higher mastery descriptions require model revision in response to evidence rather than simple accumulation of behaviours. Whether these provisions are sufficient to avoid a behaviourally shallow skill-list interpretation remains a conceptual and empirical question for Studies 1 and 6, and a negative answer would require the architecture to be reformulated.

Prior systems-thinking education in Vietnam. Vietnam already has prior experience with school-based systems-thinking education. An initiative associated with Ockie Bosch and Nam Nguyen is reported to have introduced systems thinking to Vietnamese schoolchildren approximately two decades ago, drawing substantially on system dynamics and on the Evolutionary Learning Laboratory tradition; a published account of the associated educational approach appears in Nguyen and Bosch (2014). Unpublished teaching materials associated with that work (Bosch & Nguyen, n.d.) are consistent with this characterisation: they introduce systems thinking through the iceberg model of events, patterns, systemic structures and mental models, through the contrast between linear and systemic causation, and through causal-loop modelling of a regional development case. The materials located do not appear to develop boundary critique, stakeholder legitimacy, or analysis of power and marginalisation as explicit instructional components. The available record does not establish the scale, reach, assessment arrangements, outcomes, or institutional trajectory of the initiative, and reported figures for the number of children reached are not treated here as established.

Its relevance is twofold. First, it establishes that systems-thinking education had already been introduced in Vietnamese school settings before the present proposal. Its scale, reach, pedagogical design, assessment arrangements, outcomes, and institutional trajectory remain to be established through Study 1b. Second, it raises, in the specific context this framework addresses, the questions the framework itself turns on: whether a developmental progression was specified, whether competence was assessed rather than assumed, whether learning transferred, and whether the work proved sustainable within an institutional curriculum. Systems-8 seeks to learn from and extend that experience rather than to claim a first introduction of systems thinking to Vietnam.

Table 1. Systems-8 compared with existing systems-thinking education frameworks and curriculum architectures. “Criteria addressed” refers to the five differentiation criteria stated in Section 6.2: (1) systems-thinking competence as the organising construct; (2) a multidimensional learner profile aggregating evidence across subjects; (3) critical and normative concerns operationalised as assessed competencies; (4) a specified assessment architecture for systems-thinking competence as such; and (5) transfer as a separately assessed dimension of performance. Entries marked partial indicate that a criterion is addressed in part, not that it is fully satisfied. Systems-8 specifies all five as design intentions, but none has yet been demonstrated.

Framework (source)	**Structure**	**School-age evidence**	**Validated instrument**	**Criteria addressed (full/partial)**
Systems Thinking Hierarchical model (Ben-Zvi Assaraf & Orion, 2005, 2010a, 2010b)	8 characteristics in 3 sequential levels	Grades 4 and 8; 6-year longitudinal follow-up	Qualitative elicitation battery; no closed-form psychometric instrument	1 (within earth-systems content); 4 (partial)

Framework (source)	**Structure**	**School-age evidence**	**Validated instrument**	**Criteria addressed (full/partial)**
Structure–Behavior–Function (Hmelo-Silver et al., 2007)	3 representational components	Expert–novice studies; school applications	Coding schemes; not a scored instrument	1 (partial: a representation scheme rather than a competence model)
Systems thinking skills (Richmond, 1993)	7 thinking skills	None (conceptual, adult-oriented)	None	1
Systems thinking skill set (Arnold & Wade, 2015, 2017)	8 skills plus purpose and interconnections	None (systems engineering, adults)	None	1
Habits of a Systems Thinker (Waters Center for Systems Thinking, n.d.; Benson & Marlin, 2020)	14 habits with age-differentiated materials	Extensive K-12 practitioner use	None located in peer-reviewed literature	1; 3 (partial: named as habits, not assessed)
Sustainability competence frameworks (Wiek et al., 2011; UNESCO, 2017; Bianchi et al., 2022)	Systems thinking as 1 of 5, 8, or 12 competences	Not age-specified; GreenComp is lifelong	None	3 (partial: named as competences, not assessed)
Validated school-age instrument (Mambrey et al., 2020)	3 dimensions, IRT-calibrated	Grades 5–6 (n = 196, Germany)	Yes – 36-item instrument, IRT	1; 4 (partial: an instrument, not an architecture)
DSRP (Cabrera et al., 2008)	4 cognitive patterns	Practitioner adoption in schools; claimed outcomes	None located	1; 2 (partial: domain-general, but no profile aggregation specified)
NGSS crosscutting concepts (NRC, 2012; NGSS Lead States,	7 concepts with K-12 grade-band progressions	Extensive; official standards	Assessment architecture specified (NRC, 2014); no standalone	4 (partial: specified for science proficiency, not for

Framework (source)	Structure	School-age evidence	Validated instrument	Criteria addressed (full/partial)
2013; NRC, 2014)			systems-thinking instrument	systems-thinking competence)
Australian Curriculum (ACARA)	Systems thinking in Design and Technologies F–10; Sustainability priority	Official curriculum with achievement standards	None located	3 (partial: normative concerns appear in curriculum and achievement expectations, but not as separately scored systems-thinking competencies)
IB PYP / MYP	Transdisciplinary concept-driven inquiry, ages 3–16	Extensive	None located for systems thinking	2 (partial: transdisciplinary, but not organised around a systems-thinking construct)
Systems-8 (this paper, proposed)	8 domains + 4 cross-cutting capabilities + 6 mastery levels	Not yet tested	None; validation studies proposed	All five specified as design intentions; none demonstrated

The final row should be interpreted cautiously. Several existing frameworks possess empirical, curricular, or implementation evidence that Systems-8 does not yet have, whereas Systems-8 currently contributes an integrative design proposition rather than demonstrated educational effects. The appropriate comparison at this stage is therefore between established partial contributions and an undemonstrated conjunction of design features. Whether that conjunction adds educational value is a question for the research programme in Section 15. One check on the table itself does not depend on that programme and should be carried out first. Because the scoring in Table 1 was performed by parties with an interest in its conclusions, the frameworks should be rescored against the five criteria by two coders working independently of the design team, with agreement and the substance of any disagreement reported alongside the table. Study 1 remains the systematic test of P1; this is a check on the present evidence for it.

6.2 The differentiation claim, stated narrowly

A candidate differentiation criterion rejected. Continuous age progression from approximately six to seventeen, with differentiated outcome descriptors at each band, is not treated as a differentiating feature of Systems-8. The *Next Generation Science Standards* specify grade-band progressions for each crosscutting concept across K-12, and the Australian Curriculum specifies systems thinking within Design and Technologies from Foundation to Year 10 with achievement standards at each band. Both are continuous and both are official, so age continuity cannot support a claim of architectural distinctiveness.

That comparison also clarifies the structure of the differentiation claim made here. Two requirements that might otherwise be conflated are treated separately: that systems-thinking competence is itself the object being progressed and reported, and that evidence from different subjects can contribute to a single multidimensional learner profile. The claim therefore comprises the five criteria set out below. The curriculum architectures that make this clarification necessary were located by targeted checking rather than by systematic search, and Study 1 is specified accordingly in Section 15.2.

The literature review above establishes that the following already exist: developmental hierarchies of systems-thinking characteristics (STH); analytic representation schemes (SBF); skill sets (Richmond; Arnold & Wade); a cognitive theory of the patterns underlying systems thinking (Cabrera et al.); practitioner habit frameworks with age-differentiated materials (Waters Center); normative competence statements within policy frameworks (Wiek et al.; UNESCO; GreenComp); official curriculum architectures with continuous grade-band progressions and associated assessment specifications (NGSS; Australian Curriculum); transdisciplinary conceptual curricula (IB PYP and MYP); and at least one psychometrically validated school-age instrument (Mambrey et al.).

What could not be located is a specification that simultaneously satisfies all five of the following. Each element exists somewhere; the proposition is that their integration is absent.

1. **Systems-thinking competence as the organising construct**, rather than systems concepts distributed within a subject's content structure. NGSS progresses crosscutting concepts within science and engineering; the Australian Curriculum progresses design and systems thinking within a learning area and names systems within a cross-curriculum priority. In neither case is systems-thinking competence itself the object being progressed, reported, and accumulated;
2. **A single multidimensional learner profile, rather than a composite score, aggregating evidence across subjects**, so that work in science, history, literature, civics, technology, and everyday social contexts contributes to one competency profile. This is the criterion no located framework satisfies: NGSS aggregates within science and engineering, and the Australian Curriculum reports by learning area;

Gap or theoretically justified absence? The absence of a cross-subject learner profile should not automatically be interpreted as a design omission in prior work. Fischer's (1980) skill theory provides a substantive reason to expect performance to remain context-bound, with the same learner performing differently across subject matter and support conditions. Systems-8 therefore treats cross-subject aggregation as a hypothesis rather than a defining success condition. The proposed profile is justified only if evidence from different disciplinary contexts can be interpreted on a sufficiently common scale while retaining meaningful context information. If subject context accounts for systematic variance that prevents such interpretation, evidence should remain context-tagged and criterion 2 should be withdrawn from the differentiation claim rather than narrowed to preserve it. The cost of that condition should be visible before the evidence is collected. Criterion 2 is the criterion that no located framework satisfies, so its withdrawal would not trim the differentiation claim but would leave it resting on criteria that existing curriculum architectures already address in part, and the claim would then have to be restated as something substantially weaker than a distinct architecture. Section 15.3 states this as a consequence of Gate 3. Study 5 provides the principal test of this condition, with Studies 6 and 9 examining its consequences for interpretation over time.

3. **Critical and normative systems concerns operationalised as assessed competencies** – boundary justification, purpose contestation, stakeholder legitimacy, distributional consequence – with level descriptors and scoring criteria, rather than named as curriculum values, priorities, or content. This criterion is deliberately narrow. The Australian Curriculum's Sustainability priority already addresses competing world views, equity, social justice, and intergenerational consequence. What is claimed here is therefore not the presence of normative concerns but their operationalisation as separately assessed competencies. What could not be located is any specification of what a twelve-year-old's boundary justification looks like at successive levels of attainment, or how such a judgement would be scored;
4. **A specified assessment architecture for systems-thinking competence as such**, in the sense of the National Research Council's assessment triangle (National Research Council, 2001) – a cognition model, task and artefact types, and an interpretation model. This criterion must be stated carefully. *Developing Assessments for the Next Generation Science Standards* (National Research Council, 2014) specifies exactly such an architecture for three-dimensional science learning and is the most substantial prior work of this kind. Its object, however, is science proficiency, within which crosscutting concepts are one dimension. No equivalent specification for systems-thinking competence across the curriculum could be located; and
5. **Transfer to unfamiliar contexts as a defined and separately assessed dimension of performance, rather than an assumed property of higher-order performance.**

The five elements are proposed as functionally interdependent rather than merely co-located. Treating systems-thinking competence as the organising construct provides a common object of development. A cross-subject learner profile requires comparable evidence across contexts, which in turn requires a common assessment and interpretation architecture. Critical and normative concerns are included because an architecture that excluded purpose, boundary, legitimacy, and distributional consequences would represent only a subset of the systems traditions from which it is derived. Transfer is separated because performance in familiar contexts cannot by itself establish that a cross-curricular competence has developed. The differentiation claim therefore concerns the functional dependence of these elements within one architecture, not simply their coexistence.

The differentiation claim of this paper is exactly that conjunction, and nothing broader. It is a claim about what is *specified* in the literature, not about what is *validated*: Systems-8 is itself not yet validated, as Table 1 records, and the paper does not differentiate itself on a property it does not possess. The honest comparison remains between demonstrated partial contributions and an undemonstrated integrative one. The claim is stated in a form Study 1 could falsify: if a systematic review locates a specification satisfying all five, the case for Systems-8 as a distinct standalone contribution is no longer sustained, and the appropriate response is to adopt or extend that specification rather than to publish another.

A note on nomenclature. The coincidence that the closest prior art in the education literature is also an eight-part model is unfortunate, and the numeral in "Systems-8" invites the reasonable suspicion that eight is a branding choice rather than an analytic result. Section 9.2 states the derivation and the conditions under which the number would change. The authors' present view is that academic outputs should pair the Systems-8 designation with a construct-descriptive title or subtitle rather than relying on the product name alone. The designation does not imply that eight is theoretically privileged: the number remains subject to the revision procedure in Section

9.2, and if that procedure establishes that two domains are not pedagogically distinct, the finding governs the name rather than the name constraining the finding. A structural analysis recovering fewer latent factors than there are domains is not by itself such a finding.

7. The insufficiency of direct method-by-method teaching

Responding to the methodological objection does not mean direct instruction in methodology is wrong. For older adolescents and adults, direct study of SSM, system dynamics, CSH, VSM, TOC, and critical systems practice is not only possible but desirable. Direct methodology courses preserve academic lineage, make assumptions visible, and reduce the risk of presenting a newly branded synthesis as an original systems theory. The claim is narrower: such courses are insufficient as the *sole* architecture for a long-term programme beginning in childhood, for six reasons.

First, methods are not curricula. A methodology guides inquiry or intervention within a class of situations. A curriculum must additionally specify developmental sequencing, prior knowledge, learning outcomes, scaffolding, assessment, progression, and transfer. System dynamics explains feedback and accumulation but does not specify what a seven-, ten-, thirteen-, and sixteen-year-old should each be able to do. SSM structures learning about worldviews but does not supply a cross-age competency progression. CSH exposes boundary judgements but does not define a child-accessible pathway from noticing excluded people to analysing legitimacy and power.

Second, a collection of methods is not a developmental architecture. If learners complete separate courses in SSM, system dynamics, and TOC, each may use different terminology, tasks, and assessment criteria. The programme records three completions without being able to state which common competencies developed, where the learner remains weak, or whether learning transfers. The absence of a common spine makes cumulative development difficult to represent.

Third, method-by-method teaching can produce surface tool learning. A learner may become proficient at drawing a rich picture, causal loop, stakeholder map, or current-reality tree without understanding the assumptions and limits of the representation. Tools are visible and teachable, so they can become proxies for competence. A correctly drawn diagram can rest on weak evidence, an unjustified boundary, or a misleading causal claim.

Fourth, separate methods can encourage method-driven framing. After learning causal loops, a learner may interpret every situation as a feedback problem; after TOC, every situation may appear to have one dominant constraint; after SSM, every difficulty may be framed as conflicting worldviews. This is a familiar risk in professional practice: the available method shapes what the practitioner sees. A pedagogical architecture should therefore begin with inquiry before method selection.

Fifth, method-by-method teaching does not create a shared assessment architecture. Course completion, correct terminology, and competent use of a familiar tool do not establish transfer. A learner may reproduce a causal loop in a taught environmental example but fail to recognise feedback in a family, historical, or organisational situation.

Sixth, teaching methodology directly asks young learners to handle complex ideas about how knowledge is created. The answer is not to abandon these foundational concepts for simple tools, but to translate them so they match the child's developmental level. At an early level, learners might be asked to distinguish “what I saw” from “what I think happened” before formal language about observation and inference is introduced. At a subsequent level, they might recognise that different people define success differently before encountering worldview theory. Later, they might justify why a factor is inside or outside a map before studying formal boundary critique. At an advanced adolescent level, they might compare methodologies and analyse how method choice shapes intervention. These are proposed pedagogical steps, not established developmental milestones; the age bands in Section 10 remain hypotheses to be tested.

The implication is a layered programme in which Systems-8 provides foundational inquiry habits, tool families are introduced as learners mature, method-informed courses later make theoretical origins explicit, and advanced learners study methodological choice and pluralism. Systems-8 prepares access to methodologies rather than blocking it.

This position can be stated precisely: established systems methods are not inadequate; they were developed for purposes other than serving as a unified developmental curriculum for children. Systems-8 is justified only if it provides educational functions that neither direct methodology courses nor the existing education frameworks reviewed in Section 6 provide by themselves.

8. The educational gap and Systems-8

Sections 5 to 7 identify five connected gaps. Each is stated here as a claim about the *state of the literature* rather than a claim about the merit of existing work, and each is therefore checkable.

The developmental gap. Broadly articulated progressions do exist: the NGSS grade-band progressions describe how each crosscutting concept develops across K-12, and the Australian Curriculum specifies band-level achievement standards for systems thinking within Design and Technologies. What could not be located is a progression of systems-thinking *competence as such* – one that is independent of a single learning area, that has been empirically anchored rather than expert-written, and that has been validated through construct modelling in the sense of Wilson (2005, 2009). Existing studies demonstrate that children at particular ages can engage with particular systems concepts, but the evidence is dispersed across disciplinary contexts, ages, pedagogies, and measurement approaches (Ben-Zvi Assaraf & Orion, 2005, 2010a; Feriver et al., 2019; Fisher, 2023; Hokayem et al., 2020; Mambrey et al., 2020). The STH model is the nearest approach to an empirically grounded progression and covers grades 4 to 8 within earth-systems content.

The integration gap. Traditional systems provide complementary insights, but a child-facing curriculum needs a common structure connecting them without falsely presenting them as identical. Learners require a recognisable inquiry grammar before they can make sophisticated choices among methodological traditions.

The competency gap. Many courses specify content and activities without articulating a multidimensional profile of what learners can do across contexts. A competency framework must distinguish, for example, stakeholder awareness from causal reasoning, and pattern recognition from boundary judgement. It must also accommodate uneven profiles: a learner may be strong in perspectives and weak in feedback. Notably, the frameworks with the strongest normative coverage (GreenComp, UNESCO) have the weakest measurement specification, while the framework with the strongest measurement specification (Mambrey et al., 2020) has the narrowest construct coverage.

The assessment gap. Systems-thinking competence cannot be inferred from course completion or a single diagram. Assessment needs multiple forms of evidence – explanation, transfer, model revision, delayed application – and must distinguish poor systems reasoning from insufficient subject-matter knowledge, language difficulty, or drawing skill. Under AI-enabled conditions (Section 4), this gap has become materially more consequential.

The curriculum-to-research gap. A credible programme requires a traceable chain from systems literature to pedagogical design, from pedagogy to observable artefacts, from artefacts to scoring and interpretation, and from interpretation to empirical validation. Without this chain, Systems-8 would remain a branded collection of questions rather than a defensible educational framework.

Accordingly, the primary contribution lies neither in proposing an additional universal systems methodology nor in enumerating further components of systems thinking. Rather, it offers an educational translation architecture that systematically links theoretical frameworks, developmental progression, competency models, empirical evidence, and research designs.

The hypothesised developmental mechanism is iterative rather than linear. Repeated use of a shared inquiry vocabulary across progressively varied substantive contexts is expected to help learners recognise recurring forms of systems reasoning beyond the surface features of individual tasks. Explicit justification, model revision, and comparison of alternative representations are intended to make that reasoning visible and revisable, while gradual reduction of scaffolding increases independent use. Cross-context tasks then test whether the learner can mobilise the same reasoning under changed conditions rather than merely reproduce a familiar procedure. This mechanism is a design hypothesis rather than an established account of learning; Studies 7 and 9 are intended to test whether the predicted instructional sensitivity, conditional transfer, and longitudinal development occur.

9. Introduction to Systems-8

9.1 Definition and boundaries

Systems-8 is defined provisionally as **an age-progressive pedagogical, competency, and assessment architecture that translates recurring systems-thinking concerns into a coherent inquiry structure for learners aged approximately six to seventeen.**

In this paper, systems-thinking competence is used as an organising construct family rather than as a presumption of a single latent trait. It refers provisionally to the profile formed by the core competencies and associated cross-cutting capabilities specified in Section 11. Whether those components support a higher-order general construct, several correlated dimensions, or no defensible general factor is an empirical question for Study 6 rather than an assumption built into the architecture.

The definition contains several boundaries. Systems-8 is not proposed as a new general systems theory, a superior replacement for established systems methodologies, a superior replacement for existing systems-thinking education frameworks, a universal problem-solving algorithm, or an empirically validated intervention. It should not be positioned beside SSM, system dynamics, CSH, VSM, or TOC as though it were another methodology at the same level. It is a translation layer between those foundations and educational practice (Figure 1).

Systems-8 organises inquiry into eight domains: *Situation, Purpose, People and Perspectives, Parts, Wholes and Boundaries, Patterns, Connections and Emergence, Consequences, and Change.* These domains are not claimed to be novel concepts, and Section 6 documents where each has appeared before. Their proposed contribution lies in integration, developmental translation, and linkage to competence and assessment.

Situation asks what is happening, what is observed, what is inferred, what remains unknown, and how the situation has been framed; it resists defining the problem through a preferred solution. *Purpose* asks what the system is intended to achieve, how actual outcomes differ from stated intentions, and how different actors define success. *People and Perspectives* asks who participates, decides, benefits, is affected, or remains unheard, and how differently placed actors interpret the same situation. *Parts, Wholes and Boundaries* asks what the whole consists of, what is included, excluded, nested, or connected to the wider environment, how alternative boundaries alter interpretation, and what can be said about the whole that cannot be said about any of its parts. *Patterns* asks what has changed, repeated, fluctuated, or persisted over time. *Connections and Emergence* asks how elements influence one another through direct, indirect, circular, and delayed relationships, and what system-level behaviour arises from those interactions that could not be predicted from the parts considered separately. *Consequences* asks what may occur immediately, later, elsewhere, or to groups outside the initial frame, including trade-offs and unintended effects. *Change* asks what intervention should be tried, how it will be monitored, and how action should be revised when outcomes differ from expectations.

The eight domains are not to be taught as a rigid linear checklist. Systems inquiry is recursive: a newly recognised stakeholder may change the boundary; a consequence may expose a hidden purpose; a behaviour-over-time graph may require reframing the situation; an intervention may generate a new situation requiring another cycle. The framework is a navigational structure, not an algorithm.

9.2 How the eight domains were derived – and the rule for collapsing them

A retrospective mapping of the eight domains onto established systems traditions would not by itself constitute a derivation. At a sufficient level of abstraction, almost any set of six to twelve systems concepts can be mapped onto SSM, system dynamics, and CSH. This section therefore states the derivation actually used and, more importantly, the conditions under which the domain set would change.

Derivation. The domain set was produced by a three-step design process, not by an empirical study. Step one, the recurring concerns of the major systems traditions and of the education frameworks in Section 6 were listed and compared, and concerns recurring across multiple distinct source traditions or frameworks were retained for further consideration. Step two, the retained concerns were filtered against two pedagogical criteria: whether a concrete, age-appropriate task could be written for a six-to-eight-year-old that engages a recognisable precursor of the concern, and whether the concern generates an observable learner artefact. Concerns failing either filter, such as recursion in the VSM sense or formal model calibration, were deferred to the advanced pathway rather than made foundational domains. Step three, the surviving concerns were grouped into the smallest set that preserved distinctions the design team judged pedagogically consequential, in particular the distinction between boundary judgement and situation framing, and between pattern description and causal explanation.

Step one was not a systematic review; step two used the design team's judgement rather than cognitive interviews with children; step three's "pedagogically consequential" criterion was applied by parties who are not disinterested. All three steps are exactly what Studies 1, 2, and 4 are designed to redo under conditions that permit disconfirmation.

Known overlap. Four adjacencies are theoretically weak and should be expected to blur empirically. *Situation* and *Parts, Wholes* and *Boundaries* both concern framing, and differ mainly in whether the framing question is about description or about inclusion. *Patterns* and *Connections and Emergence* correspond to what system dynamics treats as a single structure–behaviour construct, and Mambrey et al. (2020) recovered a three-dimensional rather than eight-dimensional structure in a comparable age range. *Consequences* and *Change* both concern the future, and differ mainly in whether the learner is an observer or an agent. *Purpose* and *People and Perspectives* both concern whose account of success prevails.

A further known limitation of the derivation should be recorded. It was carried out against the systems methodologies and the systems-thinking education frameworks available to the design team, and did not include the cognitive frameworks or the national curriculum architectures reviewed in Section 6. The derivation is therefore narrower than the corpus now recognised as relevant, and Study 1 is empowered to revise the domain set on that basis.

Domains, latent factors, and the pre-committed revision rule. The eight domains are hypothesised *pedagogical* distinctions. They are not a pre-commitment to eight latent psychometric factors, and the two questions should not be conflated. Structural analyses may support correlated domains, higher-order factors, a bifactor solution, or fewer latent dimensions than there are domains, and any of these outcomes is compatible with the domains remaining useful distinctions for curriculum, task design, and reporting. Factor structure is also instrument-dependent: it reflects the item set, the method, and the age range of the sample as well as the construct. Mambrey et al. (2020) recovering three dimensions establishes what their instrument measured in their sample; it does not establish how many distinctions an educational architecture should draw.

This distinction could be used to make the framework unfalsifiable, and the following procedure is stated in advance to prevent that. A failure of two domains to separate structurally does not by itself collapse them; it triggers three named tests of pedagogical distinctness, all specified here rather than after the result is known. First, a *classification test:* trained raters, given learner artefacts, assign them to one of the two domains at an acceptable level of agreement. Second, a *response-process test*: cognitive interviews show that tasks in the two domains elicit recognisably different reasoning. Third, a *differential-instruction test*: instruction targeted at one domain produces unequal gains across the two.

Where structural non-separation is accompanied by failure of all three tests, the two domains are merged, and the framework is reported at the reduced number even if this requires renaming it. Where the structural analysis does not separate the domains but the three tests distinguish them, the domains are retained and the measurement model is reported as having fewer latent dimensions than domains – an outcome that must be stated explicitly rather than presented as an eight-factor result. The adjudication is made by the independent reviewer identified in Section 15.3, not by the design team, because the team has an interest in the number being retained. The number eight is a hypothesis about pedagogical distinctions, not a commitment, and not a prediction about factor structure.

9.3 The three-layer architecture

Systems-8 is visualised across three layers (Figure 1). The first layer contains traditional systems and methodologies. The second layer contains the shared pedagogical and competency architecture. The third contains courses, tools, tasks, and assessments. A causal-loop diagram may support work in Connections and Emergence; a rich picture may support Situation and People and Perspectives; boundary questions may support Parts, Wholes and Boundaries; a current-reality tree may support Connections and Change. No tool is permanently owned by one domain, and no domain requires one fixed tool.

The central positioning statement: **Systems-8 stands neither above nor beside established methodologies; it stands between systems foundations and age-progressive educational practice.**

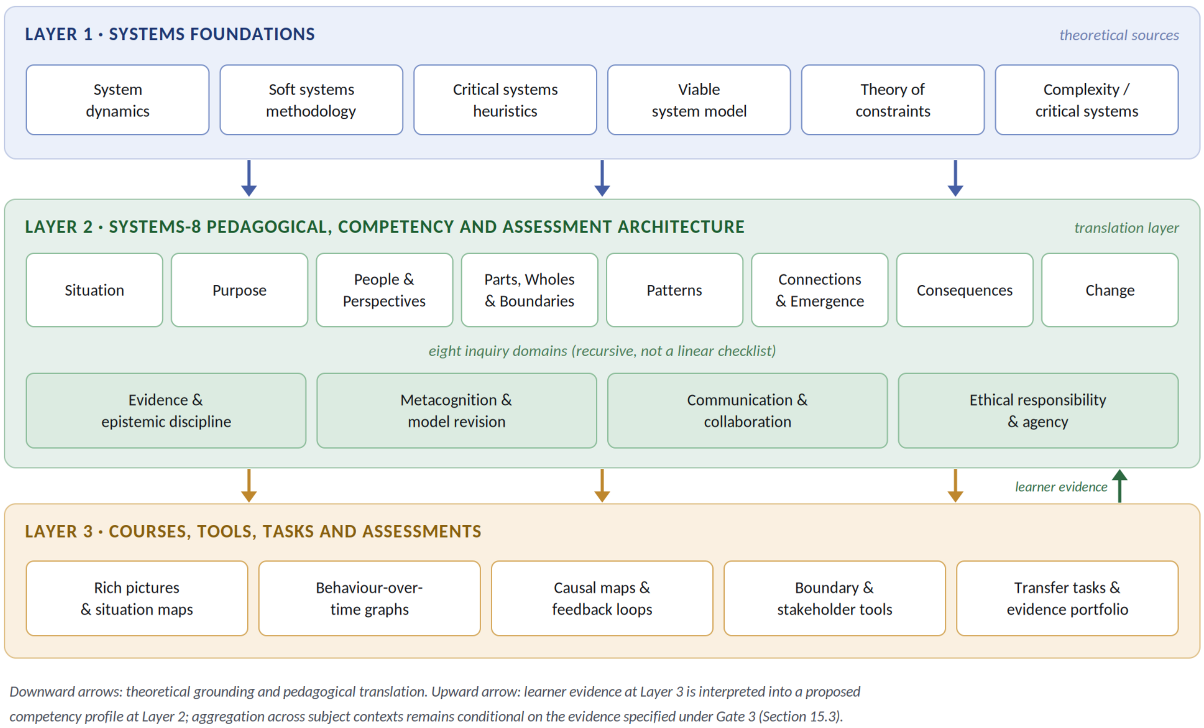

Figure 1. *The proposed three-layer architecture. Layer 1 supplies theoretical grounding; Layer 2 is the Systems-8 translation layer comprising eight inquiry domains and four cross-cutting capabilities; Layer 3 contains the courses, tools, tasks, and assessments through which learners generate evidence. Downward arrows denote theoretical grounding and pedagogical translation; learner evidence generated in Layer 3 is interpreted upward into a proposed competency profile at Layer 2; aggregation across subject contexts remains conditional on the evidence specified under Gate 3 (Section 15.3).*

9.4 Emergence as an integrative construct

An objection with force is that a framework which never requires a learner to demonstrate that system-level properties arise from interaction among parts has not, in fact, specified what distinguishes a system from a collection. Emergence is widely treated as a central characteristic of complex systems (Meadows, 2008; Verhoeff et al., 2018), and treating it only as a background concept contributed by complexity perspectives, with no position in the architecture, would leave a substantive gap rather than a gap in vocabulary.

The architecture locates *emergence* across two domains rather than adding a ninth. *Parts, Wholes* and *Boundaries* supplies the concept of the whole – the recognition that a system can be described at a level at which its parts cannot. *Connections* and *Emergence* supplies the mechanism – that interaction among parts, including feedback and organisation, produces behaviour at the level of the whole. Neither domain alone is sufficient, and the placement reflects that: a learner who has only the concept of a whole can name emergence without explaining it, and a treatment that supplies only the concept of interaction risks encouraging the misconception that a large number of connections is itself emergence.

Emergence is not proposed as a ninth domain. This is not intended to protect the number eight; Section 9.2 already treats that number as a hypothesis subject to a pre-committed revision rule. At present, there is insufficient evidence to determine whether emergence should be represented as a separable competence or as an integrative outcome across domains. The question is therefore posed for empirical resolution: does emergence appear among school-age learners as a separable competence, as an integrative outcome demonstrated across several domains, or as neither? Study 1 examines how the concept is treated in the frameworks and curricula reviewed in Section 6; Study 2 tests whether the progression stated in Section 10 can be empirically ordered; and Study 6 tests whether emergence-related items load separately or with the domains that carry them. If a separable construct is recovered, this would trigger an independently adjudicated review of the domain architecture, applying the evidential principles stated in Section 9.2, rather than automatically creating a ninth domain.

The construct also carries weight under the AI-enabled conditions described in Section 4. A generated artefact – a well-formed causal-loop diagram or a fluent explanation – can display the appearance of systemic reasoning. A polished statement about emergence is not diagnostic by itself, because an AI system can also generate such language. Evidence of emergent reasoning becomes more informative when the learner's own reasoning is made visible through unaided or supervised explanation, oral defence, revision in response to challenge, or transfer to a changed context. Section 12 therefore treats emergence-related artefacts as evidence only in combination with the conditions under which they were produced and the learner's capacity to justify them.

10. Systems-8 as a pedagogical architecture

A pedagogical architecture must explain how learning develops, not merely what topics exist. Systems-8 adopts a spiral progression in which the same domains recur with increasing abstraction, representational complexity, autonomy, epistemic responsibility, and ethical responsibility.

10.1 Four indicative age bands

The proposed six-to-eight band is labelled **Explorer** and emphasises concrete situations, whole-and-part relationships, simple sequences, immediate effects, and recognition that another person may see the same situation differently. The proposed nine-to-eleven band is labelled **Connector** and emphasises multiple causes, recurring patterns, simple feedback stories, short- and long-term effects, and missing stakeholders. The proposed twelve-to-fourteen band is labelled **System Mapper** and emphasises comparing problem frames, using behaviour-over-time representations, constructing causal and boundary maps, identifying feedback and delay, and analysing trade-offs. The proposed fifteen-to-seventeen band is labelled **System Designer** and emphasises model critique, methodological comparison, uncertainty and power analysis, and adaptive intervention design. The progression is summarised as **See, Connect, Map, Design**, and detailed in Table 2.

Table 2. Proposed age progression. The bands are design ranges for curriculum sequencing, not validated developmental stages; Section 10.2 states their status and the limits of the claim.

Age band	Stage	Progression verb	Proposed learning emphasis	Epistemic and ethical emphasis
6–8	Explorer	See	Concrete parts, people, sequence, immediate effects; another person may see this differently	Distinguishing “what I saw” from “what I think happened”
9–11	Connector	Connect	Multiple causes, recurring patterns, simple feedback stories, short- and long-term effects, missing stakeholders	Recognising that people define success differently
12–14	System Mapper	Map	Comparing problem frames; behaviour-over-time representations; causal and boundary maps; feedback, delay, trade-offs	Justifying why a factor is inside or outside the map
15–17	System Designer	Design	Model critique, methodological comparison, uncertainty and power analysis, adaptive intervention design	Analysing how method choice shapes what is seen and who is served

The emergence sequence referred to in Sections 9.4 and 11 is a developmental hypothesis derived from the design rationale in Section 9.2, not a finding of the child-development literature. Emergence is not separately assessed in the six-to-eight band, where learners work with concrete wholes and their parts and may notice that a group can do things an individual cannot; recognition is expected in the nine-to-eleven band, explanation in the twelve-to-fourteen band, and critique of emergent explanations in the fifteen-to-seventeen band. Study 2 tests whether this ordering holds.

10.2 Status and limitations of the age bands

The bands are **curriculum-design ranges aligned to common schooling structures**, not empirically derived developmental stages. This must be stated plainly, and the alignment of the bands with conventional primary/middle/secondary divisions is a design convenience that could easily be mistaken for a cognitive-developmental finding.

Three considerations bound the claim. First, the empirical literature supports *existence* rather than *sequence*: Feriver et al. (2019) at ages four to six, Ben-Zvi Assaraf and Orion (2010a) at grade four, Hokayem et al. (2020) in the elementary years, Mambrey et al. (2020) at grades five and six, and Raved and Yarden (2014) at grade seven each demonstrate that learners of that age can do particular things under particular instruction. No study establishes that these capabilities emerge in a fixed order or that earlier ones are prerequisite to later ones.

Second, developmental theory gives positive reason to expect unevenness rather than uniform stages. Fischer's (1980) skill theory holds that cognitive development consists of hierarchically constructed skills that are *context-bound*, so that performance varies substantially by context and by the degree of environmental support, producing an optimal-level/functional-level gap rather than global stage transitions. If that account is correct, a systems-thinking progression should be expected to look different in a science context and a family context for the same child on the same day. This is a strong reason to treat age bands as design ranges and to place the assessment weight on demonstrated competence.

Third, and most importantly, **the methodology for validating a claim of this kind already exists and was not used**. Learning progressions are validated through construct modelling: an explicit construct map orders qualitatively distinct levels of proficiency together with the item responses indicating each level, and is then tested with a measurement model (Wilson, 2005, 2009). The wider literature is candid that progressions are frequently proposed on plausibility and only later – sometimes never – validated (Corcoran et al., 2009; Duncan & Rivet, 2013). Study 2 of the research programme is specified accordingly: its output should be a set of **construct maps in Wilson's sense**, one per competency, empirically anchored through cognitive interviews and item response modelling, rather than a refined set of age descriptors written by experts.

Until that work is done, placement and evaluation should combine age expectations with demonstrated competence, and a younger learner demonstrating advanced stakeholder sensitivity or an older learner remaining weak in feedback reasoning should be treated as the expected case rather than the anomaly.

10.3 Cross-disciplinary application

The framework supports application across subjects, which matters because systems-thinking competence as the organising construct and cross-subject evidence aggregation are both central to the differentiation claim in Section 6.2 and because cross-disciplinary transfer cannot be assessed within a single subject area. A history case can involve evidence, competing purposes, power, causal patterns, and long-term consequences. A literature case can explore worldviews, relationships, and unintended effects. A science case can examine components, flows, feedback, and thresholds. A technology case can address purpose, data, excluded users, and governance. A family or school case can make abstract concepts concrete.

Pedagogical translation must preserve conceptual integrity: simplification should not become distortion. Young learners can explore boundary choices without formal CSH terminology, but the underlying question – who or what has been left outside the inquiry – must remain. They can recognise self-reinforcing stories before using polarity notation, but the feedback logic must remain. They can compare "success for whom?" before learning formal worldview analysis, but competing purposes must remain visible.

The long-term pathway therefore moves from Systems-8 foundations to tool families, then to method-informed courses, and eventually to methodological pluralism. At the advanced stage, learners should understand not only how to use a method but why one method may be appropriate, insufficient, or ethically problematic in a particular context.

10.4 The boundary checkpoint

The eight domains are pedagogically co-present but need not be theoretically equivalent. Boundary judgement has a recursive role, because prior boundary choices condition which purposes, people, relationships, patterns, and consequences become visible at all (Midgley, 2000; Midgley & Pinzón, 2011; Ulrich & Reynolds, 2020). Midgley and Pinzón (2011) state the dependency in the form most relevant here: value judgements about which purposes it is right to pursue and boundary judgements about what is relevant to those purposes are mutually constituting, so that a disagreement which appears to be about facts is frequently a disagreement about where the boundary of concern has been drawn. Boundary judgement is therefore revisited throughout an inquiry rather than treated once and set aside, and each domain carries an accompanying question: which situation have we chosen to look at, and what have we left outside it; whose purposes become visible within this boundary, and whose fall outside it; who becomes a stakeholder if the boundary is widened; what is inside, what is outside, and how would another boundary change what we see; does the pattern change if we look over ten years rather than one month; which relationships appear, or disappear, as the boundary moves; who outside our original frame bears the consequences; and would a different boundary lead us to choose a different intervention.

These checkpoints are pedagogical. Boundary checkpoints may be embedded across learning activities, but boundary competence is scored as a distinct target construct rather than re-scored within the rubrics of the other domains. The reason is stated in Section 12.3, and the substantive proposition to which it corresponds is stated as P11 in Section 15.1: whether boundary judgement in fact conditions the other competencies is treated as a question the evidence may answer, not as a design commitment.

11. Systems-8 as a competency framework

Systems-8 is designed around observable capability rather than content exposure alone. Each domain corresponds to a core competency, while four cross-cutting capabilities determine the quality with which the competency is enacted. Table 3 sets out domains, competency focus, and indicative evidence.

Table 3. Systems-8 domains, competency focus, and indicative evidence.

Domain	Competency focus	Indicative evidence
Situation	Evidence-based framing; observation, inference, assumption and unknown; resisting solution-led framing	Situation statement; fact–assumption–unknown table; alternative-frame comparison
Purpose	Stated and operative purposes; success for whom; conflicting goals; who holds authority to define outcomes	Purpose map; “success for whom?” analysis; multi-criteria success statement

Domain	Competency focus	Indicative evidence
People and Perspectives	Participants, decision makers, beneficiaries, affected and unheard groups; at advanced levels, power and legitimacy	Stakeholder map showing influence, impact, voice and missing perspectives; perspective-taking account
Parts, Wholes and Boundaries	Components and nested systems; system versus environment; justified inclusion and exclusion; boundary revision	Nested-system map; inside–outside map with written justification; boundary-revision note
Patterns	Sequence, repetition, trend, oscillation, threshold, delay, path dependence, multiple time horizons	Timeline; behaviour-over-time graph with explanation of the shape
Connections and Emergence	Direct, multiple, indirect, circular and delayed influence; reinforcing and balancing feedback; critique of causal models; system-level properties and behaviours arising through interaction and organisation	Causal chain; feedback narrative; causal-loop diagram with evidence annotations; explanation of a system-level behaviour naming the interactions that produce it
Consequences	Immediate, delayed, distributed, intended and unintended effects; trade-offs; risk under incomplete evidence	Consequence tree; scenario matrix; trade-off analysis naming who bears which cost
Change	Proposing an intervention; linking action to explanation; measures; scale of test; monitoring and revision	Intervention proposal with measures, review points and stated revision logic; revision log

11.1 Eight core competencies

The first, **Situation Framing**, is the capacity to describe what is happening, distinguish observation from interpretation, identify uncertainty, compare alternative frames, and avoid defining a problem through a preferred solution. The second, **Purpose Reasoning**, is the capacity to identify stated and operative purposes, compare definitions of success, recognise conflicting goals, and ask who has authority to determine desired outcomes.

The third, **Stakeholder and Perspective Awareness**, involves identifying participants, decision makers, beneficiaries, affected groups, and unheard stakeholders; understanding that perspectives differ; and, at advanced levels, analysing power, legitimacy, and representation. The fourth, **Parts, Wholes, and Boundary Judgement**, involves recognising components and nested systems, distinguishing system from environment, justifying inclusion and exclusion, revising boundaries when inquiry needs change, and articulating what can be said about the whole that cannot be said about any of its parts.

The fifth, **Pattern and Change-over-Time Reasoning**, moves from sequence and repetition to trends, oscillation, thresholds, delay, path dependence, and multiple time horizons. The sixth, Connection, Causality, Feedback, and Emergence Reasoning, moves from direct influence to multiple causation, causal chains, reinforcing and balancing feedback, indirect effects, and critique of causal models; in Mambrey et al.'s (2020) sample, distinguishing direct from indirect causal effects was a major source of difficulty. At higher levels it extends to emergent reasoning: explaining behaviour at the level of the whole that arises from interaction among parts and is not attributable to any part individually (Section 9.4). The seventh, **Consequence, Trade-off, and Uncertainty Reasoning**, includes immediate and delayed effects, intended and unintended consequences, distributional impacts, trade-offs, scenarios, risk, and incomplete evidence. The eighth, **Responsible Intervention and Adaptive Learning**, includes proposing a change, linking action to explanation, defining measures, testing at appropriate scale, monitoring side effects, and revising action.

11.2 Four cross-cutting capabilities

Evidence and Epistemic Discipline concerns the distinction among observation, inference, assumption, and claim; evaluation of source quality; and acknowledgement of uncertainty. Under AI-enabled conditions this capability carries additional weight, since it is the capability most directly implicated in evaluating generated output. **Metacognition and Model Revision** concerns explaining one's reasoning, recognising the limitations of a representation, and changing a model when evidence changes. **Communication and Collaboration** concerns explaining relationships, listening to different perspectives, building shared representations, and working constructively with disagreement. **Ethical Responsibility and Agency** concerns fairness, participation, power, responsibility, and the willingness to act while remaining accountable for consequences. It also concerns recognising when participation is not possible or not safe: ethical agency includes identifying when dialogue is blocked by coercive power or lies beyond the learner's responsibility to resolve, and knowing when adult, safeguarding, or institutional support is required rather than further inquiry. A learner who distinguishes a coercive situation from a merely contested one, and who seeks protective recourse rather than continued dialogue, is demonstrating this capability rather than failing at the others; the reasoning is set out in Section 16.

The four cross-cutting capabilities are not assumed to be latent factors embedded identically in every domain score. Their measurement status is provisional. Where they are scored, evidence should be recorded separately from the core domain score so that common scoring criteria do not artificially create covariance among the eight competencies. Study 6 should compare alternative models in which the capabilities operate as distinct constructs, cross-cutting facets, or higher-order influences, and the reporting structure should follow the model best supported by the evidence rather than presuppose twelve separable dimensions.

Learner outcomes are organised in three layers: the eight core competencies; the four cross-cutting capabilities; and transfer outcomes – the ability to mobilise these competencies in a new disciplinary or social context, select an appropriate representation, and revise the approach when context changes. Transfer is separated because success on a rehearsed task may indicate procedural familiarity rather than systems competence.

11.3 Mastery levels and their relationship to existing taxonomies

Five levels of demonstrated mastery are proposed, together with a zero level recording the absence of sufficient evidence: Not Yet Demonstrated, Recognise, Explain, Apply, Transfer, and Reflect and Adapt (Table 4). Level 0 is a statement about the evidence available rather than about the learner, and should be reported as such. In reporting to learners and families this should be expressed in those terms, for example that evidence of the competency has not yet been collected at the specified support condition this term, rather than in any wording implying that the learner has been found unable to demonstrate it. A preferred reporting phrase is “Evidence is not yet sufficient at this support condition,” rather than language implying a fixed deficit in the learner. Recognise means identifying a concept with examples or prompts; Explain means articulating it in one’s own words; Apply means using it in a familiar context; Transfer means using it in an unfamiliar context; Reflect and Adapt means critiquing assumptions, revising the model, or altering action in response to evidence. The highest level is not complexity for its own sake but disciplined revision. Mastery level and support condition are separate dimensions. Each scored observation should record whether performance was demonstrated unaided, with ordinary instructional scaffolding, or with AI assistance. An AI-assisted performance may provide useful evidence of supported capability, but it should not by itself establish the same level of unaided competence.

The placement of Transfer above Apply is itself a hypothesis. It assumes that successful use of a competency in an unfamiliar context supports a stronger evidential claim than successful use in a familiar context. Study 6 should test whether this ordering is empirically defensible. If transfer behaves primarily as a context condition rather than an ordinal proficiency level, it should be reported as a separate dimension alongside mastery and support condition rather than retained as Level 4.

It is closely related to two established taxonomies. The revised Bloom taxonomy’s cognitive-process dimension – Remember, Understand, Apply, Analyze, Evaluate, Create – covers similar ground in its lower levels, and its second, knowledge dimension (factual, conceptual, procedural, metacognitive) is directly relevant to the metacognitive emphasis of the highest Systems-8 level (Anderson & Krathwohl, 2001). The SOLO taxonomy – prestructural, unistructural, multistructural, relational, extended abstract – describes increasing structural complexity of a learner’s response and is arguably the closer analogue, since Systems-8’s progression is about relating elements rather than performing operations (Biggs & Collis, 1982).

The revised Bloom taxonomy is tabulated in Table 4 for auditability, but the correspondence at the upper levels should be treated as approximate rather than as an equivalence. Its cognitive-process dimension orders operations by kind rather than by the contextual distance at which they are performed, and the Systems-8 progression turns on precisely that distance: Apply and Transfer are the same operation under different conditions of familiarity, a distinction the Bloom ordering does not carry. The SOLO taxonomy is the closer analogue.

Two differences are claimed, and both are testable rather than asserted. First, Systems-8 requires **transfer to be evidenced separately**, and in the present mastery model this is provisionally represented as Level 4, rather than treating transfer as implicit in higher-order performance; neither Bloom nor SOLO requires evidence from an unfamiliar context to award its upper levels. Second, the top level is defined by **revision in response to disconfirming evidence** rather than by synthesis or abstraction, which reflects the specific epistemic concern of Section 4. If Study 6 finds that Systems-8 mastery ratings are empirically indistinguishable from SOLO codings of the same artefacts, the mastery scale should be replaced by SOLO rather than retained.

Table 4. Proposed mastery progression, with the nearest analogues in two established taxonomies. The analogues are approximate and are stated to make the relationship auditable, not to claim equivalence.

Level	Label	Interpretation	Nearest SOLO analogue	Nearest revised-Bloom analogue
0	Not Yet Demonstrated	No sufficient evidence at the specified support condition	Prestructural	–
1	Recognise	Identifies the feature with examples or prompts	Unistructural	Remember
2	Explain	Articulates the idea in the learner's own words	Multistructural	Understand
3	Apply	Uses the competency in a familiar context	Relational (emerging)	Apply
4	Transfer	Uses the competency in an unfamiliar context; cueing status reported separately	Relational	Analyze
5	Reflect and Adapt	Critiques assumptions, revises the model, or alters action in response to disconfirming evidence	Extended abstract	Evaluate / Create

11.4 Three reporting dimensions, not one score

Age progression, mastery progression, and support condition are separate reporting dimensions. A ten-year-old may achieve Transfer on an age-appropriate stakeholder task while a sixteen-year-old only Recognises feedback within a more complex case. The same learner may also demonstrate a competency at one level unaided and at a higher level with ordinary scaffolding or AI support. The framework should therefore produce a multidimensional profile rather than a single total score. A composite "systems index" would conceal meaningful variation and create unjustified precision, and the paper recommends against reporting one at any stage, including in commercial materials.

Cross-subject aggregation is also conditional. If Study 5 shows that subject context accounts for substantial and systematic variance such that performances cannot be interpreted on a common scale, the learner profile should remain context-tagged and evidence should not be aggregated into cross-subject competency estimates. The profile architecture must therefore earn the right to aggregate across subjects rather than assume it.

The competency profile also supports a broader learning ecosystem. Evidence from systems thinking, sustainability, AI literacy, entrepreneurship, leadership, or civic education may contribute to related capabilities, provided that mappings are theoretically justified and empirically checked rather than used to inflate a profile. The platform should be organised around learner, competency, evidence, progression, and context – not course completion alone.

12. Systems-8 as an assessment and evidence architecture

The assessment architecture is a part of the reason Systems-8 is proposed. Method-based courses can teach powerful concepts, but without a common account of cognition, evidence, and interpretation they produce disconnected results. The National Research Council's assessment triangle provides the foundation: assessment requires a model of cognition, tasks that elicit observable evidence, and an interpretation model connecting evidence to claims about learning

(National Research Council, 2001). The most substantial prior application of that triangle in an adjacent space is Developing Assessments for the Next Generation Science Standards (National Research Council, 2014), which specifies a cognition, task, and interpretation architecture for three-dimensional science learning. The object of that architecture is science proficiency, within which crosscutting concepts form one dimension; the object here is systems-thinking competence accumulated across subjects. The distinction is one of construct rather than of method, and the design decisions below draw on that precedent.

In Systems-8, the cognition component is the competency framework of Section 11; the observation component comprises artefacts and performances associated with each domain; the interpretation component comprises rubrics, mastery descriptors, rater guidance, and validity arguments. Each evidence record also carries support-condition metadata so that unaided, ordinarily scaffolded, and AI-assisted performances are not treated as equivalent observations. The architecture must be designed before large-scale claims are made.

12.1 Domain-linked evidence

Each domain generates observable evidence, summarised in the third column of Table 3. *Situation* may produce an evidence-based situation statement and a fact–assumption–unknown table. *Purpose* may produce a purpose map or a "success for whom?" analysis. *People and Perspectives* may produce a stakeholder map showing influence, impact, voice, and missing perspectives. *Parts, Wholes and Boundaries* may produce nested-system and inside–outside maps with explicit justification. *Patterns* may produce timelines and behaviour-over-time graphs. *Connections and Emergence* may produce causal chains, feedback narratives, causal-loop diagrams, and, at higher levels, an explanation of a system-level behaviour that names the interactions producing it and states why it cannot be attributed to any single part. *Consequences* may produce consequence trees, scenario matrices, and trade-off analyses. *Change* may produce a testable intervention, measures, review points, and revision logic.

12.2 Four assessment layers

The first layer is knowledge and concept recognition: short tasks establishing whether the learner recognises a stakeholder, a pattern, a delay, or a boundary. This layer is necessary but insufficient. The second is structured application, in which the learner uses prompts or templates in a familiar case. The third is transfer, in which the learner applies the competence to an unfamiliar case, ideally in another subject or social context. The fourth is reflection and adaptation, in which the learner revises a model after new evidence, explains limitations, or changes an intervention.

Assessment should use multiple response modes. Some learners reason well verbally but struggle with drawing; others construct maps fluently but cannot justify them. Evidence should therefore include written responses, diagrams, oral explanations, teacher observation, collaborative artefacts, reflections, and project outputs. There is precedent for this requirement outside the school context. In a systemic intervention developing services with young people under sixteen living on the streets of Manchester, Boyd, Brown and Midgley (2004) combined semi-structured interviews, rich pictures, a values map, and planning workshops built on a synergy of critical systems heuristics (Ulrich, 1983) and interactive planning (Ackoff, 1981). Two features of that account bear on the present argument. First, Ulrich's boundary questions were put to the young people themselves rather than reserved for the professionals: the adaptation lay in the surrounding process – separate workshops so that young people's accounts could form without competition from professional voices, mixed-gender interviewer pairs, payment for participation – rather than in withholding the concepts. Second, the choice of representation was itself informed by boundary reflection: a photograph-based elicitation method was piloted but not retained after the young people reported that the images added little to the interviews. The relevance here is specific and limited. It indicates that formal boundary questions can be made accessible to young people when the process around them is designed against the risk of marginalisation, and that the choice of representational form is a boundary judgement in its own right. It does not establish a developmental sequence, and is not cited for that purpose. Accessibility and language demands must be monitored so that a task does not unintentionally measure reading level or graphic skill more strongly than systems reasoning – a threat that Study 4 is designed to detect and Study 8 to examine across languages.

A longitudinal evidence portfolio should preserve the developmental process: initial framing, first model, feedback, revision, oral defence, transfer task, reflection, and intervention outcome. This sequence is particularly valuable under AI-enabled conditions. A final product may be heavily tool-assisted, but a trace of reasoning, revision, explanation, transfer, and the support condition under which each performance occurred provides stronger evidence of learner competence than the product alone – which is precisely the inference the evidence in Section 4 shows to be at risk.

12.3 Scoring and validity

Scoring should be analytical rather than purely holistic. A distinction must be drawn between criteria that apply across domains and criteria that belong to the domain whose construct they define. Relevance, coherence, use of evidence, and revision apply generally. Boundary justification, relational and causal reasoning, and consequence awareness do not: each is the defining content of a particular competency, and scoring it within the rubrics of other domains would distribute a single construct across the whole instrument. Complexity should not be rewarded automatically: a map with twenty arrows is not better than one with six well-supported relationships. Nor should all perspectives be treated as equally factual; respect for perspective diversity does not remove the need for evidence.

Boundary judgement in teaching and measurement. Boundary judgement occupies an unusual position in the architecture, and its treatment differs deliberately between pedagogy and assessment. Pedagogically it is pervasive: Section 10.4 sets out the recursive checkpoint through which it recurs in every domain.

For measurement the treatment is different, and the reason should be stated rather than left implicit. If a scored boundary element were embedded in the tasks of all eight domains, boundary-related variance would be present in every score the architecture collects. Two consequences follow, both undesirable. The internal structure tested under P7 would carry a method factor introduced by design rather than by the construct, making any failure to separate uninterpretable. More seriously, the claim that boundary judgement is associated with performance in the other domains – stated as P11 – would become untestable, since every result would exhibit the predicted pattern whether or not the claim were true.

Boundary judgement is therefore assessed within its own domain, and the tasks associated with the other seven domains do not carry boundary elements as scored components. The architecture treats boundary primacy as a proposition the evidence may support or refute, not as a design commitment. A practical scoring rule follows from this distinction. A Connections and Emergence task may prompt a learner to reconsider missing influences, but its rubric should score causal and relational reasoning rather than award points for boundary justification. A rubric that awards boundary points within that task would violate the separation rule and contaminate the test of P11.

Validity must be understood as an argument supporting particular interpretations and uses of scores, not as a property conferred once on an instrument (AERA, APA, & NCME, 2014; Kane, 2013; Messick, 1995). Systems-8 assessment will require evidence regarding content representation, response processes, internal structure, relations with other variables, consequences of use, rater consistency, developmental sensitivity, and transfer. A small pilot showing pre-post improvement cannot validate the architecture, and should not be described as doing so.

12.4 Feasibility and resource requirements

The architecture above is considerably more expensive than conventional assessment, and a proposal that does not say so is not usable. Rich evidence demands teacher time for task administration, artefact review, oral questioning, rater calibration, and portfolio maintenance. If every competency were assessed at every layer for every learner every term, the workload would be prohibitive in most school settings, and the framework would fail in implementation regardless of its conceptual merit.

Table 5 therefore distinguishes a **core evidence set** that is intended to be sustainable in ordinary classroom conditions from an **extended set** used for research, moderation, or high-stakes reporting. Four principles govern the reduction. *Sampling*: not every competency needs evidence every term; a rotating sample maintains profile currency at a fraction of the cost. *Embedding*: evidence should be drawn from work learners were going to do anyway rather than added as separate assessment events. *Calibrated exemplars*: rater agreement is cheaper to sustain through annotated exemplars than through repeated moderation meetings. *Staged intensity*: oral defence and delayed transfer tasks – the most informative and most expensive elements – should be reserved for one or two points per year per learner rather than applied continuously.

AI may assist with administrative aspects such as collation, transcription, and provisional suggestions, but scoring decisions that contribute to a learner's profile should remain under human review, and the use of AI in scoring must itself be reported as part of the validity argument rather than treated as an implementation detail.

Table 5. Core and extended evidence sets. Time estimates are design targets to be tested in Study 7's feasibility phase, not measured values; a framework that cannot meet them should be simplified rather than exhorted.

The per-learner figures in Table 5 should be read at class level before they are read as feasible. On the core set, assuming three terms and a class of thirty learners, the design targets sum to approximately fifty-two minutes per learner per year, or roughly twenty-six hours per class per year, with a further one to two hours for rater calibration. For a primary teacher responsible for a single class this is close to an hour of the working week across the school year, which is demanding but not implausible. For a secondary subject teacher carrying four or five classes the same targets imply a multiple of that figure, and the core set would then require further sampling, a reduced number of competencies assessed per teacher, or distribution of the evidence load across the subject teachers contributing to a learner profile. P12 tests enactability against these limits, so the limits must be stated per teacher and per timetable rather than per learner. The figures are design targets, not measured values, and Study 7 should report observed rather than intended workload.

Evidence type	Core set (intended to be sustainable)	Extended set (research, moderation, high-stakes)	Design target, teacher time
Domain artefacts	Rotating sample: 2–3 domains per term, embedded in existing subject work	All eight domains per term, purpose-built tasks	≤ 10 min per learner per term (core)
Structured application	Template-supported task in a familiar case, one per term	Multiple cases across two or more subjects	Embedded in existing marking
Transfer task	One unfamiliar-context task per year	Cued and uncued transfer tasks, plus delayed follow-up	≈ 15 min per learner per year
Oral explanation	One short oral defence per year, may be conducted in small groups	Individual recorded oral defence per assessed competency	≈ 5–10 min per learner per year
Revision evidence	Revision log attached to one existing project	Full model-revision trace with rater commentary	Embedded
Rater calibration	Annotated exemplars, asynchronous	Formal moderation meetings, double-marking, reliability estimation	≈ 1–2 hr per teacher per year (core)
Portfolio maintenance	Automated collation; teacher confirms rather than compiles	Curated portfolio with contextual and support-condition metadata for research use	Platform-dependent; to be measured in feasibility testing
Indicative annual total (core set)	≈52 min per learner per year across the rows above	Not estimated; scales with competencies, cases, and moderation	≈26 h per class of 30 per year, plus 1–2 h calibration; multiply by classes taught

12.5 Reporting

Reporting should serve different stakeholders without collapsing evidence into surveillance. Learners should see strengths, emerging capacities, evidence, and next steps. Parents should receive meaningful developmental narratives rather than notifications of minor errors. Teachers should see cohort patterns, common misconceptions, and instructional needs. Researchers should retain de-identified artefacts, scoring decisions, contextual variables, support-condition metadata, and uncertainty estimates. Commercial partners should receive cohort-level or implementation-level summaries without identifiable child data; “aggregated” in this context does not authorise a composite systems score for an individual learner. Where substantial AI or other support was used, reporting should keep supported and unaided evidence distinguishable.

The proposed assessment architecture is what distinguishes Systems-8 from a sequence of child-friendly questions. It creates a traceable connection from inquiry to competence, from competence to evidence, and from evidence to researchable claims.

12.6 Safeguarding and sensitive evidence

Some Systems-8 tasks, particularly those concerning power, exclusion, coercion, boundary judgement, and ethical agency, may elicit disclosures about real experiences rather than hypothetical reasoning. Such disclosures must not be treated as ordinary scored evidence. Implementation protocols should distinguish assessment from safeguarding response, provide age-appropriate routes for referral to trained adults under applicable institutional and legal requirements, and minimise the inclusion of sensitive personal material in longitudinal portfolios. Where sensitive information must be retained for safeguarding purposes, it should be stored separately from competency evidence, with restricted access, defined retention rules, and clear procedures for informing learners and caregivers subject to safeguarding obligations. Research and platform design should follow data-minimisation principles and the child-protection and privacy requirements of each implementation jurisdiction. Sensitive safeguarding material should not be included in datasets released for external scrutiny merely because identifiers have been removed.

13. The transfer assumption and its critics

A central part of the proposed architecture depends on an assumption that the paper has so far stated rather than defended: that systems-thinking competence developed in one disciplinary context can be mobilised in another, and that cross-subject evidence can therefore contribute to a coherent competency profile. Transfer is provisionally the fourth mastery level and a principal outcome of Study 7; it is also examined through Studies 4 and 9, and it provides part of the rationale for cross-subject evidence aggregation. A large and serious literature holds that assumptions of this kind are usually wrong. This section states that literature at its strongest, because a framework whose transfer assumption is contested should say so in the body of the paper rather than in a limitations note.

13.1 The case against

Far transfer is rare and shrinks under rigorous designs. Barnett and Ceci (2002) showed that the transfer literature's apparent contradictions arise from failing to specify what kind of transfer is at issue, and set out nine dimensions of content and context along which transfer varies; on their mapping, the cells most relevant to educational claims are the least evidenced. Sala et al. (2019), in a second-order meta-analysis aggregating ten meta-analyses, 332 effect sizes, and 21,968 participants, found that far-transfer effects approach zero once publication bias and active control conditions are accounted for, across training types and populations, concluding that the absence of skill generalisation is "an invariant of human cognition." Sala and Gobet (2017, 2019) reach the same conclusion from chess, music, and working-memory training.

Generic thinking skills may not be teachable as such. Willingham (2008) argues that critical thinking is not a transferable general skill trainable independently of content, because thinking is structured around surface features until deep domain knowledge is available; students who learn a critical-thinking move in one domain typically fail to recognise when to deploy it elsewhere. Tricot and Sweller (2014) go further, arguing that generic-cognitive skills are biologically primary and acquired without instruction, so teaching them yields no measurable gain, and that apparent successes reduce on inspection to domain-specific knowledge acquisition. Kirschner et al. (2006) add that minimally guided inquiry – a common vehicle for teaching such skills – is poorly matched to novice cognitive architecture.

Developmental theory predicts unevenness rather than a general capacity. Fischer's (1980) skill theory, discussed in Section 10.2, holds that skills are context-bound control structures, so a "level" of systems thinking should be expected to vary by domain and support condition rather than to characterise the learner.

The systems-thinking literature itself supplies a cautionary case. Booth Sweeney and Sterman (2000) found that stock-flow performance among highly educated adults did not vary systematically with prior education in mathematics or science. This does not show that mathematical or scientific education caused no improvement, nor does it test explicit instruction in stock-flow reasoning. It does show that prior quantitative education was not a reliable indicator of performance on these accumulation tasks in their sample. That makes it unsafe to assume that a systems capability acquired or practised in one context will automatically appear in another.

13.2 Boundaries of the claim: what Systems-8 does and does not claim

Taken together, this literature does not permit the claim that Systems-8 will produce a general systems-thinking ability that learners deploy across contexts. That claim would be inconsistent with the best available evidence, and the framework does not make it.

Three narrower positions remain defensible, and the paper commits to these.

First, the framework's target is conditional transfer. The transfer literature is most pessimistic about *spontaneous* far transfer – deploying a capability in a new context without cueing. It is considerably less pessimistic about transfer under conditions of varied practice across multiple disciplinary contexts, explicit abstraction, and cueing, which is what Perkins and Salomon (1992) term high-road transfer through mindful abstraction, and what Salomon and Perkins (1989) analyse as its mechanism. Systems-8's design features that bear on this – deliberate use of multiple disciplinary contexts (Section 10.3), explicit naming of the inquiry domains as a shared vocabulary, and required reflection on why a representation was chosen – are precisely the conditions under which the literature finds transfer least unlikely. The framework should therefore be understood as a *bet on the conditions for high-road transfer*, and Study 7 should be designed to test cued and uncued transfer separately, reporting both.

Second, task-specific disciplinary knowledge is a prerequisite, not a confound to be controlled away. Willingham's and Tricot and Sweller's arguments should be read as design constraints rather than objections to be rebutted. Systems reasoning cannot be assessed in the absence of content the learner knows, and a low Systems-8 score in an unfamiliar context may indicate missing subject knowledge rather than missing systems competence. Study 6 accordingly measures or controls task-specific disciplinary knowledge; Study 4 constructs transfer tasks in contexts where prior knowledge can be independently established; and no interpretation should attribute low performance to systems competence without evidence about content familiarity.

Third, incremental validity is the decisive empirical value test, but failure is possible. If Systems-8 competency scores do not add explanatory value beyond measured critical thinking, general reasoning ability, and task-specific disciplinary knowledge on independent criterion outcomes, then the framework adds explanatory burden without explanatory value and its principal justification fails. This is stated as Proposition P8 in Section 15 and is the most consequential empirical value test in the programme.

13.3 Rationale and justification for testing

Three considerations justify continued testing of the proposal, and readers should judge whether they are sufficient. The first requires the paper to correct a defence it might otherwise be tempted to make.

It would be convenient to argue that the transfer literature's most negative findings concern brief, decontextualised training aimed at a narrow cognitive faculty and therefore say little about multi-year, content-rich, cross-disciplinary education. Booth Sweeney and Sterman (2000) provide an important caution against that move: stock-flow performance among their highly educated participants did not vary systematically with prior mathematics or science education. Lengthy quantitative education therefore cannot be assumed to generate this particular form of reasoning incidentally. The study does not, however, test whether explicit instruction in stock-flow reasoning is effective.

The distinction that matters is therefore between incidental acquisition and explicit instruction. Booth Sweeney and Sterman (2000) did not examine participants' prior curricula in sufficient detail to establish whether stock-flow reasoning had been explicitly named, practised across varied contexts, or assessed. Their result therefore cannot support the claim that mathematics and science education does not teach stock-flow reasoning. It supports the narrower claim that prior education alone was not a reliable indicator of performance in their sample. Systems-8 must test whether explicit and varied practice improves the targeted reasoning and whether any gains transfer.

This narrows what the framework may claim. The relevant question is not whether long education transfers, but whether reasoning that is explicitly named, deliberately varied, and practised across disciplinary contexts transfers under cueing. These are the high-road conditions described by Perkins and Salomon (1992) and analysed by Salomon and Perkins (1989). That literature is less unfavourable than the far-transfer meta-analyses, but it is not favourable, and P9 is stated so that it can fail.

Systems-8 rests on a contested assumption, adopts the design conditions most favourable to it, and is instrumented so that the assumption can be disconfirmed.

14. Relationship between Systems-8 and established systems methodologies

Systems-8 maintains explicit traceability to its foundations, both to prevent unacknowledged appropriation and to make advanced pathways visible.

Situation is informed by SSM's attention to problematic situations and by critical inquiry into framing. *Purpose* draws from SSM's purposeful activity, VSM's concern with identity and policy, and CSH's examination of motivation and legitimacy. *People and Perspectives* draws from SSM worldviews and CSH's stakeholder and witness categories. *Parts, Wholes and Boundaries* draws from general systems concepts, CSH, systemic intervention, and recursive organisational thinking. *Patterns* and *Connections and Emergence* draw strongly from system dynamics – behaviour over time, endogenous causality, feedback, accumulation, delay – and from Richmond's (1993) articulation of dynamic and closed-loop thinking. *Consequences* draws from system dynamics, CSH, complexity, and critical systems thinking, because effects may be delayed, distributed, contested, and boundary dependent. *Change* draws from all traditions but differently: learning-oriented accommodation in SSM, leverage in system dynamics, boundary critique in CSH, viability in VSM, focused improvement in TOC, and pluralist intervention in critical systems practice.

This mapping is conceptual, not empirical, and it is confirmatory in form. It demonstrates that the domains have recognisable ancestry; it does not demonstrate that they are the right decomposition, and Section 9.2 states why. It should not imply that any methodology reduces to one or two domains, nor that Systems-8 contains the depth of any methodology.

A traceability statement that records only what was carried across is incomplete, and the omissions are as consequential as the inclusions. Three bodies of content in the source methodologies are not represented in the domain set. Recursion in the viable system model sense – the property that a viable system contains and is contained by viable systems at other levels – was judged not to admit an age-appropriate task for a six-to-eight-year-old and was deferred to the advanced pathway. Formal model calibration and quantitative simulation in system dynamics were deferred on the same grounds. The distinction between first-order and second-order cybernetic positions is not developed at all.

Each deferral has a cost. Without recursion, learners may treat "the system" as a single fixed level rather than as a choice of level. Without calibration, a learner may take a qualitative causal map to be a model in the sense that a modeller would mean. These are consequences of a boundary judgement made by the design team, and are recorded here so that the judgement is visible and contestable – which is the standard the framework itself applies to inquiry.

Relationship to DSRP. DSRP (Section 6.1) is a close conceptual comparator rather than a further methodological source. It is not among the traditions from which the domains were derived, and Systems-8 does not claim to improve on it as an account of the patterns underlying cognition. The relationship is one of different objects, as Section 6.1 sets out.

The status of boundary judgement. Boundary judgement occupies two positions in this architecture that should not be conflated. It is one assessable competency, corresponding to one domain. It is also, on an argument advanced within the critical systems tradition, a condition on inquiry as a whole, since prior boundary choices determine which purposes, actors, relationships, and consequences become available for consideration at all. This paper adopts the first as a design decision and treats the second as a proposition (P11) rather than building it into the instrument. Section 12.3 states why the distinction matters for measurement.

The analogy of a foundational grammar is useful but limited. Systems-8 may function like a grammar of systems inquiry, providing recurring questions that organise thought before learners specialise. Yet methodologies are not merely instruments or styles; they contain distinct philosophical commitments, and advanced education must return to those commitments explicitly.

A staged pathway is proposed. Stage one develops Systems-8 foundations. Stage two introduces tool families – rich pictures, behaviour-over-time graphs, causal maps, boundary questions, constraint logic, simple simulations. Stage three introduces method-informed courses with clear attribution and theoretical context. Stage four develops methodological choice and pluralism: which approach is appropriate, what it excludes, and how methods may be combined responsibly. The existence of this pathway addresses the concern that Systems-8 may dilute established traditions. The framework should be judged partly by whether it prepares learners to understand those traditions more deeply.

15. Propositions and research programme

Systems-8 is a conceptual proposition. Its claims are separated here into numbered propositions so that each can be tested, and so that the framework's overall status can be read from which propositions currently have support. Table 6 maps propositions to studies.

15.1 Twelve propositions

P1 (Distinctiveness). No located systems-thinking framework or curriculum architecture simultaneously specifies all five criteria in Section 6.2. Falsified if a systematic review identifies one that does. P1 tests external distinctiveness only; it does not establish that Systems-8 itself has operationalised the five criteria to an adequate standard. That internal specification depends on successful construct, task, scoring, and validity work in Studies 2–6. Until then, the five criteria describe an intended architecture rather than a demonstrated contribution.

P2 (Foundational construct coverage). The eight domains cover, without major omission or redundancy, the recurring systems concerns judged suitable for foundational pedagogical translation within the proposed six-to-seventeen architecture. Advanced concepts explicitly deferred by the pedagogical filters in Section 9.2, including formal model calibration, VSM recursion, and first-order/second-order cybernetic distinctions, lie outside the scope of this proposition. Falsified if Study 1 or expert review identifies major recurring age-translatable concerns absent from the set, or if redundancy between domains is confirmed through the structural and pedagogical distinctness procedure specified in Section 9.2.

P3 (Developmental plausibility). For each competency, qualitatively distinct levels of proficiency can be described that are ordered, observable, and interpretable by teachers, and that learners at the nominated age bands can be located on. *Falsified if* construct maps cannot be constructed or do not order empirically.

P4 (Comprehensibility). Learners at each age band interpret the domain language and task prompts as intended. *Falsified if* cognitive interviews show systematic misinterpretation not remediable by rewording.

P5 (Content validity). Independent experts rate the competencies and descriptors as relevant, clear, distinguishable, observable, age-appropriate, and cross-culturally interpretable and appropriate at conventional thresholds. Falsified if content validity indices fall below accepted criteria. P5 is a design-stage content and interpretability check; it does not substitute for the later empirical cross-cultural and measurement-equivalence test in P10, and support for P5 should not be treated as evidence that P10 has been satisfied.

P6 (Scorability). Trained raters using analytic rubrics achieve acceptable inter-rater and intra-rater agreement on learner artefacts. *Falsified if* agreement remains below conventional thresholds after calibration.

P7 (Internal structure). The competency set shows an interpretable internal structure. *Note:* this proposition is deliberately not stated as "eight factors will be recovered," and recovery of fewer latent dimensions than there are domains – as Mambrey et al. (2020) found in a comparable age range – does not by itself falsify the domain set. The revision procedure in Section 9.2 applies: structural non-separation triggers the three tests of pedagogical distinctness, and only their joint failure requires the domains to be merged. A higher-order or bifactor solution is an acceptable outcome and must be reported as such rather than described as an eight-factor structure. *Falsified if* the score structure is unstable across plausible specifications or samples, or if no model supports sufficiently reliable and interpretable distinctions for the intended reporting uses.

P8 (Incremental validity). Systems-8 competency scores explain incremental variance in independently scored criterion outcomes requiring systems reasoning, such as unfamiliar-context reasoning, model revision, or intervention judgement, beyond measured critical thinking, general reasoning, task-specific disciplinary knowledge, and parsimonious rival representations of systems-thinking competence specified independently before confirmatory Study 6 analyses. Criterion outcomes used for P8 must not be the same tasks used to construct the Systems-8 competency scores. Rival models derived from Study 1 should be preregistered before Study 6 data are analysed. Reduced models suggested by Study 6 may also be examined, but comparisons arising from the same data are exploratory unless evaluated through sample splitting, cross-validation, or an independent replication sample. Among the empirical value propositions, P8 is the decisive test: failure removes the principal empirical justification for maintaining Systems-8 as a separate competency and assessment architecture.

Quantitative decision rules. No quantitative proposition should be treated as unsupported solely because a conventional null-hypothesis test is non-significant. Before confirmatory testing of P6 to P12, wherever a quantitative decision rule applies, the relevant preregistration should specify the smallest effect size or difference of educational interest, the decision procedure used when absence or practical equivalence is substantively interpreted, the required statistical power or corresponding Bayesian evidence threshold, and the sample-size assumptions implied by the final instrument and analysis model. For P8 in particular, failure requires evidence that incremental value falls below the preregistered threshold, not merely failure to reject a zero-effect null hypothesis. The same requirement governs P9 and P11, where a non-significant difference or a non-significant association is not by itself evidence that the predicted effect is absent, and P12, whose reliability, fidelity, and workload thresholds are preregistered rather than judged after the fact. Numerical thresholds are not fixed in this conceptual paper because the instruments, reliabilities, parameter counts, and criterion outcomes required to justify them do not yet exist; specifying them is therefore a prerequisite for confirmatory testing rather than a post hoc choice.

P9 (Conditional transfer). Learners taught through the framework across multiple disciplinary contexts demonstrate cued transfer to unfamiliar contexts at rates exceeding comparison conditions; uncued transfer is reported separately and is not predicted to be large (Section 13). *Falsified if* cued transfer shows no advantage over comparison conditions at or above the preregistered smallest difference of educational interest, evaluated under the decision procedure stated above; a non-significant difference alone is not sufficient.

P10 (Cross-cultural interpretability). Domain concepts retain intended meaning across the initial target contexts, and measurement invariance holds at a defensible level. *Falsified if* concepts such as stakeholder voice, legitimate disagreement, or individual agency are systematically reinterpreted in ways that alter what is measured.

P11 (Boundary gateway relation). Independently measured boundary-judgement competence is associated with performance across other Systems-8 competencies beyond age, measured task-specific disciplinary knowledge, and general reasoning and, where P7 supports an interpretable general Systems-8 factor, after accounting for that factor. Falsified if the association falls below the preregistered smallest effect of interest once the applicable controls are included, evaluated under the decision procedure stated above; a non-significant association is not by itself evidence of absence. The pattern and relative strength of the associations across Systems-8 domains is examined exploratorily and is not predicted in advance: the argument from which the proposition derives holds that prior boundary choices condition which purposes, actors, relationships, and consequences become visible at all, which does not license a prediction that some domains are less affected than others. One qualification is stated in advance because it bears on interpretation rather than on direction: an association of uniform strength across the seven other competencies is more consistent with a general factor than with a gateway relation, and should be reported as such.

Pre-committed consequence, stated in two stages. An association does not license a structural change, because it is equally consistent with boundary competence being a consequence of general systems capability, or with both arising from a common cause. The commitment is therefore staged. If P11 is not supported, boundary judgement remains one assessable domain among the set, and the pervasive pedagogical treatment described in Section 10.4 is reviewed for proportionality; this conclusion is acted on directly. If P11 is supported, the architecture is not revised on that basis: support triggers a further pre-committed study asking whether strengthening boundary-judgement competence produces improvement in the other competencies, and only the outcome of that study licenses giving boundary judgement a superordinate position, with domain descriptors, progression, and reporting model altered accordingly.

P12 (Enactability). After a feasible programme of professional development, practising teachers can administer the core assessment architecture, distinguish pedagogical prompts from scored constructs, apply the boundary-scoring separation rule, and achieve acceptable scoring reliability and implementation fidelity within the workload limits specified in Table 5. Falsified if ordinary classroom teachers, after the defined training, cannot reach preregistered reliability and fidelity thresholds or cannot enact the core architecture within feasible workload limits. Studies 5 and 7 provide the principal tests.

Relationship between P1 and P8. P1 and P8 play different gatekeeping roles. Failure of P1 removes the claim of architectural distinctiveness. Failure of P8 removes the principal empirical justification for maintaining Systems-8 as a separate competency and assessment architecture. Among the empirical propositions, P8 is therefore the decisive value test. A null result on P8 can remove the case for maintaining a separate architecture, whereas a positive result licenses only the next stage of testing.

Table 6. Propositions, corresponding studies, and current evidential status. The status column is the register referred to in Section 15.3 and should be updated publicly as studies report.

Proposition	Claim in brief	Principal study	Status (August 2026)
P1	No existing framework satisfies all five criteria of 6.2; Systems-8 internal specification remains to be demonstrated	Study 1	External distinctiveness untested; systematic review in Study 1 pending. Internal specification incomplete pending Studies 2–6.

Proposition	Claim in brief	Principal study	Status (August 2026)
P2	The eight domains cover recurring age-translatable foundational systems concerns without major gap or redundancy	Studies 1, 3	Untested; known overlap documented in 9.2; derivation did not include cognitive frameworks or curriculum architectures
P3	Ordered, observable proficiency levels can be described per competency	Studies 2, 9	Untested; no construct maps yet exist
P4	Learners interpret domain language and prompts as intended	Studies 2, 4	Untested
P5	Experts rate descriptors as valid at conventional thresholds	Study 3	Untested
P6	Raters achieve acceptable agreement under preregistered decision rules	Study 5	Untested
P7	The competency set shows an interpretable internal structure	Study 6	Untested; collapse rule pre-committed (9.2)
P8	Incremental validity beyond generic abilities, disciplinary knowledge, and preregistered rival systems-thinking models	Study 6	Untested – decisive empirical value test; equivalence/power decision rule to be preregistered before confirmatory testing
P9	Cued transfer exceeds comparison conditions	Studies 4, 7, 9	Untested; prior literature is unfavourable (13); effect and decision thresholds to be preregistered
P10	Concepts retain meaning across target cultures	Study 8	Untested; design-stage interpretability work under P5 does not substitute for P10
P11	Boundary-judgement competence associated with other	Study 6	Untested; instrumentation kept

Proposition	Claim in brief	Principal study	Status (August 2026)
	competencies beyond specified controls		clean to preserve testability (12.3)
P12	Practising teachers can enact and score the core architecture reliably after feasible training	Studies 5, 7	Untested; training, reliability, fidelity, and workload thresholds to be preregistered

15.2 Nine initial studies

The programme begins with nine planned studies. Additional studies may be triggered by pre-committed decision rules, including the causal boundary study specified under P11.

Study 1 – Systematic conceptual review and prior-implementation analysis (tests P1, P2). This study has two workstreams with different methods.

Workstream 1a – conceptual review. A systematic rather than purposive review comparing definitions and dimensions of systems thinking across education, system dynamics, SSM, CSH, VSM, TOC, complexity, sustainability education, and critical systems practice, including the frameworks and curriculum architectures in Section 6 and non-English-language sources. It should establish construct boundaries and identify which dimensions are recurring, method-specific, or absent. Its scope explicitly includes cognitive frameworks of systems thinking outside the education literature, national and international curriculum architectures with continuous progressions, Vietnamese-language sources and programme records, and measurement instruments rather than frameworks alone. At least two reviewers should independently code candidate frameworks against the five differentiation criteria, with agreement and disagreements reported rather than resolved only by the design team.

Workstream 1b – historical and implementation analysis. A documentary and interview-based study of the prior Vietnamese initiative described in Section 6.1, examining its scale, pedagogical design, age differentiation, assessment arrangements, evidence of outcomes, institutional trajectory, and the reasons for its continuation or discontinuation. Sources include publications, programme archives, training materials, funder and institutional records, and interviews with those who designed and delivered it. This workstream addresses a question the published literature cannot answer, since institutional discontinuation is rarely documented in the research record. It involves human participants and therefore requires ethics approval in its own right, which the conceptual review does not. Where those who designed the programme participate as informants or co-authors, they should not also serve as its independent evaluators; Section 15.3 states the separation of roles.

Because the review conducted for the present paper was carried out by parties with an interest in its conclusions, and because two categorical omissions have already been identified, Study 1 should be led or independently audited by a researcher without a stake in the outcome.

Sequencing note. Study 1 logically precedes the specification in Sections 9 to 12, and the present paper reverses that order. This is a real weakness, mitigated in two ways: the framework is labelled v0.3 and designated a hypothesised specification (Section 9.2), and Study 1 is empowered to alter the domain set, with the collapse rule stated in advance. If Study 1 is conducted by the same team that produced the specification, an independent reviewer should adjudicate any decision to retain a domain the review does not support.

Study 2 – Construct maps and developmental progression (tests P3, P4). A review of developmental and educational research combined with a modified Delphi process involving systems scholars, child-development specialists, teachers, curriculum designers, and assessment experts. Its output should be the construct maps in Wilson's (2005, 2009) sense – ordered, empirically anchored proficiency levels paired with the item responses indicating each level – not expert-written age descriptors. Cognitive interviews with learners at each band should test whether language and tasks are understood as intended. Following task refinement, pilot response data should be collected across the age bands and analysed using an appropriate item-response or other measurement model to test whether the proposed ordering is empirically recoverable. Participatory co-design should involve learners from the initial target contexts in reviewing or generating task forms, especially for perspective, agency, boundary judgement, and ethical responsibility, so that cognitive interviews do not remain the only form of learner participation.

Study 3 – Content validation (tests P5). Experts rate each competency and descriptor for relevance, clarity, distinctiveness, observability, age appropriateness, and cross-cultural interpretability and appropriateness. Item- and scale-level content validity indices may be combined with qualitative analysis of disagreement. Items with apparent consensus should still be examined for disciplinary bias, and the expert panel should include scholars who are sceptical of domain-general competency frameworks. The expert panel should include representation from the initial target cultural contexts so that construct definition is not fixed within one cultural boundary before Study 8.

Study 4 – Task development (tests P4, and prepares P9). Scenario-based tasks across family, school, history, literature, environment, science, technology, and society. Think-aloud protocols and cognitive interviews should reveal whether learners use the intended reasoning or rely on irrelevant cues. Alternative response formats should be tested to reduce language and drawing bias. Transfer tasks should be constructed in contexts where prior content knowledge can be independently established (Section 13.2). Participatory task co-design with learners should complement think-aloud and cognitive-interview procedures, particularly for tasks involving perspective, agency, boundary judgement, and ethical responsibility.

Study 5 – Reliability, generalisability, and enactability (tests P6, P12). Research raters and practising classroom teachers, after a feasible specified training programme, independently score the same artefacts using analytic rubrics; inter-rater and intra-rater reliability are estimated separately for each group. The study should test whether teachers can distinguish pedagogical prompts from scored constructs, including the boundary-scoring separation rule in Section 12.3, and should record training time and recurrent scoring workload. With sufficient samples, generalisability theory can estimate the relative contributions of learner, task, subject context, rater type, and rater. This matters because systems performance may vary substantially across contexts – which is what Fischer's (1980) account predicts – and because an assessment architecture that only expert research raters can use reliably is not educationally enactable.

Study 6 – Construct and incremental validity (tests P7, P8, P11). Relations among the core competencies and cross-cutting capabilities are tested. The framework should show convergent relations with related constructs such as critical thinking and metacognition while retaining discriminant value. Task-specific disciplinary knowledge must be measured or controlled. P8 is tested against independently scored criterion outcomes requiring systems reasoning, such as unfamiliar-context reasoning, model revision, or intervention judgement; these criterion tasks must be distinct from the tasks used to construct the Systems-8 competency scores. The pre-committed revision rule (Section 9.2) governs the structural analysis, and the mastery-scale comparison against SOLO codings (Section 11.3) is conducted here. The structural analysis additionally tests P11 by estimating whether independently measured boundary-judgement competence is associated with the remaining competencies after accounting for a general factor, age, measured task-specific disciplinary knowledge, and general reasoning. The pattern and relative strength of these associations across competencies are examined exploratorily rather than predicted in advance. It also tests whether emergence-related items load separately or with the domains that carry them (Section 9.4). Competing established or parsimonious systems-thinking models identified through Study 1 should be preregistered before Study 6 data are analysed and compared directly with the full Systems-8 specification so that incremental value is not judged only against generic cognitive abilities. Reduced models suggested by Study 6 itself should be treated as exploratory unless evaluated through sample splitting, cross-validation, or an independent replication sample. Alternative measurement models should also test whether the four cross-cutting capabilities function as distinct constructs, cross-cutting facets, or higher-order influences.

Study 7 – Sensitivity to instruction, transfer, and classroom enactment (tests P9, P12). Pilot programmes use baseline, post-programme, cued transfer, uncued transfer, and delayed tasks. Initial studies may focus on feasibility, acceptability, implementation fidelity, and preliminary learning signals; later studies should use comparison or wait-list groups where ethical and practical. Given the evidence in Section 4, post-tests contributing to learning claims should be proctored, and delayed unaided performance should be reported alongside assisted performance. Support condition should be recorded for every scored observation so that AI-assisted, ordinarily scaffolded, and unaided performances remain distinguishable. Claims must remain proportional to design quality. Teacher preparation, fidelity, workload, and scoring reliability are outcomes rather than merely background conditions: Study 7 should test whether practising teachers can sustain the core architecture after feasible professional development and should identify which implementation failures require simplification before effectiveness claims are pursued.

Study 8 – Cross-cultural and linguistic validity (tests P10). Initial contexts may include Vietnam, Australia, and New Zealand. Concepts such as authority, participation, individual agency, stakeholder voice, and legitimate disagreement may be interpreted differently. Translation must address conceptual equivalence rather than word substitution. With adequate samples, measurement invariance can be examined.

Study 9 – Longitudinal development (tests P3, P9 over time). The strongest claim of an age-progressive architecture requires evidence across time: whether early competencies predict later performance, whether profiles remain uneven, how transfer develops, and how methodology-specific learning interacts with the common spine. Ben-Zvi Assaraf and Orion's (2010b) six-year follow-up is the closest existing precedent and a reasonable design reference.

15.3 Stage-gates, claim governance, and research independence

Stage-gates and decision rules. The research programme should not be interpreted as a sequence that automatically proceeds because an earlier proposition has merely avoided rejection. Gate 1 follows Study 1: if prior art removes architectural distinctiveness, the programme shifts from validating a standalone framework to adopting, extending, or integrating the stronger specification. Gate 2 follows Studies 2–4: if construct maps cannot be ordered empirically or tasks cannot elicit the intended reasoning, scoring and validation studies should pause while the architecture is revised. Gate 3 follows Study 5: progression to full construct and incremental-validity testing requires acceptable preregistered reliability, generalisability, teacher enactability, and workload evidence. Gate 3 also carries a consequence for the differentiation claim rather than for measurement alone. Cross-subject aggregation is a condition of criterion 2 in Section 6.2, and criterion 2 is the criterion that no located framework satisfies. If the generalisability analysis shows that subject context accounts for systematic variance preventing interpretation on a sufficiently common scale, criterion 2 is withdrawn rather than relaxed, the learner profile remains context-tagged, and the differentiation claim in Section 6.2 must be restated accordingly. Such a finding has the same standing as a Gate 1 finding that prior art already satisfies the criterion set, and should be reported as a change to what the framework claims rather than as a measurement problem to be engineered around. Gate 4 is the empirical value gate in Study 6: if P8 falls below its preregistered practical-value threshold, the principal case for a standalone competency and assessment architecture is removed. Gate 5 follows Studies 7 and 8: transfer and cross-cultural claims are restricted to the contexts in which the corresponding evidence is supported. Before Gate 1, public claims should remain conceptual only; before Gates 2 and 3, claims should be limited to design and feasibility; effectiveness claims require evidence from appropriately controlled instructional studies; and cross-context or cross-cultural claims require the corresponding transfer and invariance evidence. A proposition that cannot be tested because an earlier gate failed should be recorded as not reached, not as surviving by default.

Cybee Research Lab may develop commercial educational products based on Systems-8, creating a potential conflict between evidential and commercial interests. The safeguards should therefore be structural: preregistration of confirmatory studies; peer-reviewed publication of findings, including null or unfavourable results; transparent version control of the framework; a public register recording the current evidential status of each proposition; and access to de-identified research materials for external scrutiny where ethically permitted, excluding safeguarding material and other sensitive evidence that should not be released merely because identifiers have been removed. Where a study bears directly on a proposition whose failure would materially affect the framework's justification, particularly P8, independent replication should be sought rather than treated as optional.

A further separation is required where the research programme examines prior work whose originators are also advisors to the present project. Three roles should be held by different parties: programme insiders or co-researchers, who provide primary material and contextual interpretation; independent methodological auditors or advisors, who assess the quality of the evidence; and an adjudicating panel, which determines how the evidence should affect the framework. An advisor who designed a programme may participate in the first role but should not occupy the second or third role in relation to that programme. Research cycles may overlap in practice, but claims should not precede the evidence needed to support them.

16. Critical challenges and safeguards

Rebranding. Reviewers may argue that Systems-8 renames existing concepts. The safeguard is explicit attribution, the comparative analysis in Section 6, and a differentiation claim (Section 6.2) narrow enough to be falsified by Study 1. The framework's proposed novelty is not the invention of situation, purpose, stakeholders, boundaries, patterns, feedback, consequences, or intervention. The comparison includes DSRP, whose four patterns overlap with four of the eight domains, together with the curriculum architectures reviewed in Section 6.1. The differentiation claim is deliberately narrow: age continuity is not treated as a differentiator, and distinctiveness rests only on the five-part conjunction specified in Section 6.2.

Eclecticism. Combining elements from multiple traditions can produce an incoherent collection. The safeguard is conceptual analysis of compatibility and difference: Systems-8 must identify where traditions align, where they conflict, and which distinctions are deliberately preserved (Section 14).

Oversimplification. Eight domains can become a checklist. The safeguard is recursive design, explicit model revision, and tasks requiring learners to revisit earlier domains. Completion of all eight headings must never be treated as proof of quality, and scoring rubrics should not reward coverage per se.

Loss of methodological depth. Learners may remain permanently at an introductory synthesis. The safeguard is the staged pathway into tool families, method-informed courses, and methodological pluralism (Section 14). Systems-8 should be a bridge, not a ceiling.

False universality. Not every inquiry requires equal attention to all eight domains, and some situations demand specialised analysis. The safeguard is to teach purposeful selection and proportionality.

Dialogue is not always sufficient. The framework's emphasis on perspectives, participation, and boundary critique carries a risk of implying that conflicts dissolve once enough viewpoints are included and enough dialogue takes place. Midgley (1997) sets out the limiting case explicitly: where coercion has produced a closure of debate – where a powerful party declines to participate, excludes other voices, or where no forum for contestation exists – boundary questioning loses its purchase. Midgley's argument is that critical systems heuristics presupposes the possibility of communication, or of arbitration where communication breaks down, and therefore "becomes redundant when coercion is experienced." His conclusion is that coercion can be addressed adequately only by widening systems practice to include campaigning and direct political action – that is, by going beyond the method rather than applying it harder.

This has a specific consequence for a framework addressed to children. Bullying, abuse of authority, authoritarian family or institutional structures, and systematic exclusion are situations in which a child taught that careful listening and reasoned articulation resolve disagreement may conclude, when they do not, that the failure is their own. Midgley's own response for adult contexts – political action and campaigning – does not translate directly. The principle translates as three requirements: that learners are taught the limits of dialogue alongside its methods; that responsibility for resolving coercion is never placed on the child; and that recognising when adult, safeguarding, or institutional intervention is required is treated as part of the competence rather than as a failure of it. The corresponding capability statement is in Section 11.2. Stated positively, the two halves of this position come from the same body of work and should be taught together. Midgley and Pinzón (2011) argue that boundary critique has implications for conflict prevention as well as resolution – that supporting people to examine differences in their boundaries of concern before positions entrench can defuse a developing conflict, and that some boundary models help in recognising the systemic conditions that produce stereotyping, stigmatisation, and the victimisation of minorities. Midgley (1997) marks the limit of the same approach once debate has closed. A defensible pedagogy therefore teaches boundary reflection early, as a means of understanding difference before it hardens, and teaches alongside it the recognition that some situations have passed the point at which inquiry is the appropriate response. The operational safeguarding implications of such tasks are addressed in Section 12.6.

Construct overlap. Situation framing, consequence reasoning, metacognition, and intervention overlap with critical thinking, problem solving, design thinking, and scientific inquiry. The safeguard is a clear nomological network and the incremental-validity test of P8. Systems-8 should not claim exclusive ownership of shared capabilities. A specific instance is boundary judgement, which the architecture treats as pervasive in teaching and targeted in measurement; the safeguard against contaminating the whole instrument with a single construct is stated in Section 12.3, and the proposition it preserves is P11.

Disciplinary-knowledge dependence. Systems reasoning cannot operate in a vacuum, and the literature in Section 13 indicates this is a first-order constraint rather than a caveat. The safeguard is to distinguish enabling foundations from target competencies, use multiple disciplinary contexts, measure content familiarity, and refuse interpretations that attribute low performance to systems competence without such evidence.

Transfer failure. The framework may simply not produce transfer. The safeguard is that transfer is separately assessed, that cued and uncued transfer are reported separately, and that P9's failure is treated as a finding rather than an implementation problem.

Assessment burden. Rich evidence demands substantial teacher time. The safeguard is the core/extended distinction in Table 5, sampling of competencies, digital portfolio support, calibrated exemplars, and restriction of AI to administrative assistance under human review.

Commercial influence. A framework owned by an education venture may be promoted before validation. The safeguard is the governance in Section 15.3, together with the public claim register and clear labelling of conceptual, pilot, and validated claims. Public and commercial material describing Systems-8 must carry an equivalent status disclosure, which need not reproduce this paper's version designation but must convey three things: that the architecture is proposed rather than validated; that no claim of instructional effectiveness or measurement quality is made; and that a research programme is in progress, with a reference through which its current status can be checked. Readers should treat any Systems-8 material that does not convey all three as not authorised by this paper.

These challenges do not merely qualify the proposal; they define the conditions under which Systems-8 would deserve academic and educational legitimacy.

17. Discussion

The conceptual contribution of Systems-8 lies in separating four levels that are often conflated. A systems *methodology* guides inquiry or intervention. A systems *pedagogy* organises learning experiences. A systems *competency framework* specifies what learners can do. A systems *assessment architecture* defines what evidence supports interpretations of learner development. Established methodologies are rich at the first level and contribute to the others. Existing education frameworks and curriculum architectures vary across the second, third, and fourth levels. Few located approaches specify an assessment architecture for systems-thinking competence as such. NGSS is an important adjacent exception, but its assessment object is three-dimensional science proficiency rather than cross-curricular systems-thinking competence. The proposal is that this construct-specific assessment layer, integrated with pedagogy and competency development, remains comparatively under-specified.

The pedagogical contribution is a common spine across age groups and courses. Without such a spine, an education ecosystem becomes a catalogue of methods and activities. With it, a learner's work in a history case, an environmental simulation, a family problem, and an AI ethics project can contribute to a coherent profile – *provided* that the mappings are valid and the evidence comparable, which is an empirical question and not a design assumption.

The assessment contribution is at least as important. Systems thinking is often praised broadly and measured narrowly: existing measures focus on causal maps, feedback reasoning, or domain-specific modelling, and the strongest validated school-age instrument recovers three dimensions within one subject (Mambrey et al., 2020). Systems-8 proposes a broader architecture spanning framing, purpose, perspectives, boundaries, patterns, causality, consequences, intervention, evidence, metacognition, collaboration, and ethics. That breadth creates measurement difficulty and increases the risk that the construct will not hold together – a risk the paper accepts and instruments against rather than minimising.

The treatment of age continuity illustrates the claim-governance principle built into the framework. Targeted checking of prior art showed that continuous age progression is already well represented in the Next Generation Science Standards and the Australian Curriculum, so age continuity is excluded from the differentiation claim rather than defended as a novel feature. The same principle is formalised prospectively in Section 15: evidence that undermines a claim is intended to change the architecture or its positioning rather than to be treated as an implementation inconvenience.

The framework may also contribute to AI-era assessment. When final outputs can be generated or heavily assisted, education needs evidence of reasoning processes, transfer, revision, oral explanation, and the conditions under which performance was produced. The evidence in Section 4, including substantial self-reported AI misuse in one higher-education study, very low detected AI-related malpractice in a different school-examination context, and a causal study showing a divergence between assisted practice performance and subsequent unaided performance, indicates that conventional product-only inference warrants increasing caution. Systems-8 therefore combines artefacts with explanation, revision, transfer, and support-condition metadata to make the evidence chain more visible. This does not eliminate misuse or measurement error, but it supports a more defensible basis for interpreting learner performance.

Methodological pluralism remains the long-term goal. The framework should not train learners to apply eight headings mechanically. It should prepare them to ask which systems concern is salient, what method may help, what the method excludes, and what ethical responsibilities accompany intervention. At advanced levels, the ability to refuse or revise a method may matter as much as the ability to apply one.

Systems-8 also has implications for platform architecture. A conventional learning management system is organised around users, courses, enrolments, completions, and scores. A competency-and-evidence ecosystem must additionally represent competencies, descriptors, artefacts, contexts, raters, revisions, transfer, and uncertainty. The educational model should shape the data model rather than being retrofitted after course delivery.

The proposal is justified conditionally. It must demonstrate conceptual distinctiveness, developmental plausibility, teachability, enactability, assessment validity, transfer, cultural adaptability, and incremental value. If research shows that the eight domains are redundant, that the age progression is unsupported, that Systems-8 competencies are inseparable from existing constructs, or that a simpler existing framework – the STH model, the Habits, or a sustainability competence framework with added assessment – performs equally well, then Systems-8 should be refined, consolidated with the stronger existing framework, or repositioned so that its role reflects the evidence. A credible research programme must remain open to substantive revision when its core propositions are not supported, and Section 15 identifies P8 as the proposition most central to the framework's continuing justification.

The two central objections can now be answered together. *Why not build courses directly from established systems methodologies?* Because those methodologies are not interchangeable curriculum units and were not designed collectively as a developmental, competency, and assessment system for children. *Why not adopt an existing systems-thinking education framework?* Because each supplies part of what is needed – a developmental hierarchy, a representation scheme, a skill list, a set of habits, a normative competence statement, a validated instrument – and none supplies the conjunction specified in Section 6.2. Direct methodology courses and existing frameworks should both remain part of the pathway. The claim is that a common educational architecture is additionally required to sequence learning, connect outcomes, assess transfer, and represent development across methods and contexts – and that this claim is now stated in a form that evidence can defeat.

18. Conclusion

Artificial intelligence does not reduce the importance of knowledge. It increases the importance of knowing enough to evaluate generated outputs and of possessing the judgement to decide what should be done with them. Evidence now shows that AI-assisted performance can diverge from subsequent unaided performance, while product-only assessment becomes harder to interpret when a tool can generate polished outputs. The six risks in Section 4 are design concerns of different evidential strength: construct-validity concerns are immediate, whereas developmental effects of sustained AI substitution remain to be tested. These conditions strengthen the case for assessment that distinguishes supported performance from unaided competence and for education that keeps reasoning, judgement, and agency visible.

Systems thinking offers relevant capabilities because it directs attention to purposes, relationships, boundaries, patterns, feedback, perspectives, consequences, and adaptation. This relevance does not by itself establish a case for Systems-8. Established systems methodologies, systems-thinking education frameworks, and curriculum architectures already provide substantial theoretical, pedagogical, developmental, and assessment resources. The unresolved claim identified in this paper is narrower: no located architecture simultaneously specifies systems-thinking competence as the organising construct, a multidimensional learner profile aggregating evidence across subjects, critical and normative systems concerns as assessed competencies, an assessment architecture for systems-thinking competence as such, and transfer to unfamiliar contexts as a separately assessed dimension of performance.

Systems-8 is not proposed as a competing systems methodology nor as a replacement for those existing approaches, but as an integrated pedagogical, competency, and assessment architecture. Its eight domains provide a common inquiry spine. Its competency framework defines observable capabilities. Its proposed mastery architecture provisionally distinguishes recognition, explanation, application, transfer, and adaptive revision, while support condition is recorded separately so that AI-assisted and unaided performances are not conflated. Its assessment architecture links cognition, artefacts, interpretation, and longitudinal evidence. Its research programme subjects twelve numbered propositions to conceptual and empirical scrutiny and specifies in advance the conditions under which the architecture should be refined, consolidated, repositioned, or integrated with stronger existing approaches.

A central assumption of the architecture is that explicitly taught systems capabilities can be mobilised across disciplinary contexts under specified conditions. A substantial transfer literature contests broad claims of this kind. Systems-8 therefore does not assume a generic systems-thinking ability. It targets conditional transfer, tests cued and uncued transfer separately, treats task-specific disciplinary knowledge as a prerequisite, and makes incremental validity the decisive empirical value test while Study 1 separately tests architectural distinctiveness.

The central proposition is consequently both ambitious and restrained: foundational knowledge remains essential, while explicit systems-thinking capabilities may help learners connect and use that knowledge when situations are interconnected, dynamic, and contested. Systems-8 is a proposed means of making those capabilities teachable, observable, assessable, and their transfer explicitly testable. It warrants adoption only if research demonstrates that this integration adds educational value without erasing methodological depth, and only if the research is conducted under conditions that would permit an alternative conclusion to be reached and published.

Declarations

Authors and affiliations. James Nguyen, Cybee Research Lab, Cybee Learning, Melbourne, Victoria, Australia. Dr. Tinh Doan, Cybee Research Lab. Dr. Ha Tran, Cybee Research Lab.

Correspondence. Dr. Tinh Doan's email: thanhtinhdoan@gmail.com; Dr. Ha Tran's email: ha.tranlethu@gmail.com; and James Nguyen's email: ngocson2004@gmail.com. Cybee Research Lab, Melbourne, Victoria, Australia.

Funding. This work was funded by Cybee Research Lab.

Competing interests. Cybee Research Lab is developing, or may develop, a commercial educational product based on the framework described in this paper. The authors therefore have a financial interest in the framework's adoption. Section 15.3 states the governance safeguards intended to separate evidence from commercial interest. Readers should weigh the arguments in this paper with that interest in view. Any additional role conflicts arising in Study 1b or later empirical studies, including advisory or prior-programme relationships relevant to the evidence under review, should be disclosed in the corresponding empirical publication.

Ethics approval. Not applicable to this conceptual paper, which reports no research involving human participants. Ethics approval will be sought for all studies involving human participants, including Workstream 1b and Studies 2–9 as applicable. Study 1a is a literature-based conceptual review and does not involve human participants. The approving body should be named in the resulting empirical publications.

Data availability. No datasets were generated or analysed. Materials associated with the framework – domain descriptors, draft rubrics, and task specifications – will be made available for external scrutiny under the terms described in Section 15.3. Sensitive safeguarding material and other data whose release could create a foreseeable risk to a child will not be included in external research packages merely because direct identifiers have been removed.

Use of generative AI. Generative AI tools were used as editorial support for language refinement, clarity, and manuscript structure. The conceptual arguments, interpretation of the literature, source verification, and final decisions were undertaken by the authors, who take full responsibility for the manuscript.

References

Ackoff, R. L. (1981). *Creating the corporate future*. Wiley.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete ed.). Longman.

Arnold, R. D., & Wade, J. P. (2015). A definition of systems thinking: A systems approach. *Procedia Computer Science, 44*, 669–678. https://doi.org/10.1016/j.procs.2015.03.050

Arnold, R. D., & Wade, J. P. (2017). A complete set of systems thinking skills. *INSIGHT, 20*(3), 9–17. https://doi.org/10.1002/inst.12159

Australian Curriculum, Assessment and Reporting Authority. (n.d.). *Australian Curriculum: Technologies*; *Sustainability cross-curriculum priority*. ACARA.

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. https://doi.org/10.1037/0033-2909.128.4.612

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. https://doi.org/10.1073/pnas.2422633122 (See also the published correction: *PNAS, 122*(34), e2518204122. https://doi.org/10.1073/pnas.2518204122)

Beer, S. (1985). *Diagnosing the system for organizations*. John Wiley & Sons.

Benson, T., & Marlin, S. (2020). *The habit-forming guide to becoming a systems thinker* (2nd ed.). Waters Center for Systems Thinking.

Ben-Zvi Assaraf, O., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching, 42*(5), 518–560. https://doi.org/10.1002/tea.20061

Ben-Zvi Assaraf, O., & Orion, N. (2010a). System thinking skills at the elementary school level. *Journal of Research in Science Teaching, 47*(5), 540–563. https://doi.org/10.1002/tea.20351

Ben-Zvi Assaraf, O., & Orion, N. (2010b). Four case studies, six years later: Developing system thinking skills in junior high school and sustaining them over time. *Journal of Research in Science Teaching, 47*(10), 1253–1280. https://doi.org/10.1002/tea.20383

Bianchi, G., Pisiotis, U., & Cabrera Giraldez, M. (2022). *GreenComp: The European sustainability competence framework* (Y. Punie & M. Bacigalupo, Eds.; EUR 30955 EN). Publications Office of the European Union. https://doi.org/10.2760/13286

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J.

R. Pomerantz (Eds.), Psychology and the real world: Essays illustrating fundamental contributions to society (pp. 56–64). Worth Publishers.

Booth Sweeney, L., & Sterman, J. D. (2000). Bathtub dynamics: Initial results of a systems thinking inventory. *System Dynamics Review, 16*(4), 249–286. https://doi.org/10.1002/sdr.198

Bosch, O. J. H., & Nguyen, N. C. (n.d.). *Introduction to systems thinking* [Unpublished teaching materials, Module 1].

Boyd, A., Brown, M., & Midgley, G. (2004). Systemic intervention for community OR: Developing services with young people (under 16) living on the streets. In G. Midgley & A. E. Ochoa-Arias (Eds.), *Community operational research: OR and systems thinking for community development* (pp. 203–252). Kluwer Academic/Plenum Publishers.

Cabrera, D. (2006). *Systems thinking* [Doctoral dissertation, Cornell University].

Cabrera, D., & Colosi, L. (2008). Distinctions, systems, relationships, and perspectives (DSRP): A theory of thinking and of things. *Evaluation and Program Planning*, 31(3), 311–317. https://doi.org/10.1016/j.evalprogplan.2008.04.001

Cabrera, D., Colosi, L., & Lobdell, C. (2008). Systems thinking. *Evaluation and Program Planning*, 31(3), 299–310. https://doi.org/10.1016/j.evalprogplan.2007.12.001

Campbell, D. T. (1979). Assessing the impact of planned social change. Evaluation and Program Planning, 2(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Checkland, P., & Poulter, J. (2020). Soft systems methodology. In M. Reynolds & S. Holwell (Eds.), *Systems approaches to making change: A practical guide* (2nd ed., pp. 201–253). Springer. https://doi.org/10.1007/978-1-4471-7472-1_5

Chirikov, I., Smirnov, I., & Kizilcec, R. F. (2026). Generative AI use and misuse call for assessment reform in higher education. *Science, 392*(6800), 818–820. https://doi.org/10.1126/science.aec5115

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report No. RR-63). Consortium for Policy Research in Education.

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Duncan, R. G., & Rivet, A. E. (2013). Science learning progressions. *Science, 339*(6118), 396–397. https://doi.org/10.1126/science.1228692

Feriver, Ş., Olgan, R., Teksöz, G., & Barth, M. (2019). Systems thinking skills of preschool children in early childhood education contexts from Turkey and Germany. *Sustainability, 11*(5), 1478. https://doi.org/10.3390/su11051478

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477–531. https://doi.org/10.1037/0033-295X.87.6.477

Fisher, D. M. (2023). Systems thinking activities used in K-12 for up to two decades. *Frontiers in Education, 8*, 1059733. https://doi.org/10.3389/feduc.2023.1059733

Forrester, J. W. (1994). System dynamics, systems thinking, and soft OR. *System Dynamics Review, 10*(2–3), 245–256. https://doi.org/10.1002/sdr.4260100211

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006 (See also the published correction: *Societies, 15*(9), 252. https://doi.org/10.3390/soc15090252)

Gmyrek, P., Berg, J., Kamiński, K., Konopczyński, F., Ładna, A., Nafradi, B., Rosłaniec, K., & Troszyński, M. (2025). *Generative AI and jobs: A refined global index of occupational exposure* (ILO Working Paper No. 140). International Labour Organization. https://doi.org/10.54394/HETP0387

Goldratt, E. M. (1990). *What is this thing called theory of constraints and how should it be implemented?* North River Press.

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert–novice understanding of complex systems. *Journal of the Learning Sciences, 16*(3), 307–331. https://doi.org/10.1080/10508400701413401

Hokayem, H., Jin, H., & Yamaguchi, E. (2020). Feedback loop reasoning and knowledge sources for elementary students in three countries. *Eurasia Journal of Mathematics, Science and Technology Education, 16*(2), em1819. https://doi.org/10.29333/ejmste/112582

International Baccalaureate Organization. (n.d.). *Primary Years Programme*; *Middle Years Programme*. IBO.

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review, 10*(1), 18–26. https://doi.org/10.1007/s13162-020-00161-0

Jackson, M. C. (2019). *Critical systems thinking and the management of complexity: Responsible leadership for a complex world*. John Wiley & Sons.

Jackson, M. C. (2024). *Critical systems thinking: A practitioner's guide*. John Wiley & Sons. https://doi.org/10.1002/9781394203604

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* (arXiv:2506.08872) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2506.08872

Lehmann, M., Cornelius, P. B., & Sting, F. J. (2025). *AI meets the classroom: When do large language models harm learning?* (arXiv:2409.09047) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.09047

MacInnis, D. J. (2011). A framework for conceptual contributions in marketing. *Journal of Marketing, 75*(4), 136–154. https://doi.org/10.1509/jmkg.75.4.136

Mambrey, S., Timm, J., Landskron, J. J., & Schmiemann, P. (2020). The impact of system specifics on systems thinking. *Journal of Research in Science Teaching, 57*(10), 1632–1651. https://doi.org/10.1002/tea.21649

Meadows, D. H. (2008). *Thinking in systems: A primer* (D. Wright, Ed.). Chelsea Green Publishing.

Merola, R., Ernst, E., Samaan, D., del Rio-Chanona, M., & Teutloff, O. (2026). *Workers' exposure to AI: What indicators tell us – and what they don't* (ILO Brief No. 1). International Labour Organization. https://doi.org/10.54394/00033279

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Miao, F., Shiohira, K., & Lao, N. (2024). *AI competency framework for students*. UNESCO. https://doi.org/10.54675/JKJB9835

Midgley, G. (1997). Dealing with coercion: Critical systems heuristics and beyond. *Systems Practice*, 10(1), 37–57.

Midgley, G. (2000). *Systemic intervention: Philosophy, methodology, and practice*. Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-1-4615-4201-8

Midgley, G. (2008). Response to paper "Systems thinking" by D. Cabrera et al.: The unification of systems thinking: Is there gold at the end of the rainbow? *Evaluation and Program Planning*, 31(3), 317–321. https://doi.org/10.1016/j.evalprogplan.2008.04.002

Midgley, G., & Pinzón, L. A. (2011). Boundary critique and its implications for conflict prevention. *Journal of the Operational Research Society*, 62(8), 1543–1554. https://doi.org/10.1057/jors.2010.76

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment* (J. W. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). National Academies Press. https://doi.org/10.17226/10019

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. National Academies Press.

Nguyen, N. C., & Bosch, O. J. H. (2014). The art of interconnected thinking: Starting with the young. *Challenges*, 5(2), 239–259.

OECD. (2019a). *OECD learning compass 2030: A series of concept notes*. OECD Publishing.

OECD. (2019b). *Knowledge for 2030* (OECD Future of Education and Skills 2030 concept note). OECD Publishing.

OECD. (2019c). *Student agency for 2030* (OECD Future of Education and Skills 2030 concept note). OECD Publishing.

Ofqual. (2025). *Malpractice in GCSE, AS and A level: Summer 2025 exam series* [Official statistics]. Office of Qualifications and Examinations Regulation.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., Vol. 11, pp. 6452–6457). Pergamon.

Raved, L., & Yarden, A. (2014). Developing seventh grade students' systems thinking skills in the context of the human circulatory system. *Frontiers in Public Health, 2*, 260. https://doi.org/10.3389/fpubh.2014.00260

Richmond, B. (1993). Systems thinking: Critical thinking skills for the 1990s and beyond. *System Dynamics Review, 9*(2), 113–133. https://doi.org/10.1002/sdr.4260090203

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4, 155–169. https://doi.org/10.1007/BF01405730

Sala, G., Aksayli, N. D., Tatlidil, K. S., Gondo, Y., & Gobet, F. (2019). Near and far transfer in cognitive training: A second-order meta-analysis. *Collabra: Psychology, 5*(1), Article 18. https://doi.org/10.1525/collabra.194

Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science, 26*(6), 515–520. https://doi.org/10.1177/0963721417712760

Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in Cognitive Sciences, 23*(1), 9–20. https://doi.org/10.1016/j.tics.2018.10.004

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist, 24*(2), 113–142. https://doi.org/10.1207/s15326985ep2402_1

Sterman, J. D. (1994). Learning in and about complex systems. *System Dynamics Review, 10*(2–3), 291–330. https://doi.org/10.1002/sdr.4260100214

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Irwin/McGraw-Hill.

Tricot, A., & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review, 26*(2), 265–283. https://doi.org/10.1007/s10648-013-9243-1

Ulrich, W. (1983). *Critical heuristics of social planning: A new approach to practical philosophy*. Haupt.

Ulrich, W., & Reynolds, M. (2020). Critical systems heuristics: The idea and practice of boundary critique. In M. Reynolds & S. Holwell (Eds.), *Systems approaches to making change: A practical guide* (2nd ed., pp. 255–306). Springer. https://doi.org/10.1007/978-1-4471-7472-1_6

UNESCO. (2017). *Education for Sustainable Development Goals: Learning objectives*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000247444

Vennix, J. A. M. (1996). *Group model building: Facilitating team learning using system dynamics*. John Wiley & Sons.

Verhoeff, R. P., Knippels, M.-C. P. J., Gilissen, M. G. R., & Boersma, K. T. (2018). The theoretical nature of systems thinking: Perspectives on systems thinking in biology education. *Frontiers in Education, 3*, Article 40. https://doi.org/10.3389/feduc.2018.00040

Waters Center for Systems Thinking. (n.d.). *Habits of a systems thinker*. Retrieved August 2026, from https://waterscenterst.org/

Wiek, A., Withycombe, L., & Redman, C. L. (2011). Key competencies in sustainability: A reference framework for academic program development. *Sustainability Science, 6*(2), 203–218. https://doi.org/10.1007/s11625-011-0132-6

Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review, 109*(4), 21–32. https://doi.org/10.3200/AEPR.109.4.21-32

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730. https://doi.org/10.1002/tea.20318

World Economic Forum. (2025). *The future of jobs report 2025*. World Economic Forum.