How Do We Learn from System Dynamics Models?

A Structured Review across Engagement Modes, Conceptualizations of Learning, and Measurement Approaches

Prepared for the 44th International Conference of the System Dynamics Society, 2026

Authors: Ellen O'Neill[1], David Keith[2], Nici Zimmermann[3], Hazhir Rahmandad[4], Saras Chung[5]

Abstract

System dynamics rests on a premise that is rarely tested directly: that engaging with a model improves people's mental models and, through them, their reasoning and decisions about complex systems. We review 32 empirical studies (2010–2024) that examine some version of this premise and organize them along three interrelated dimensions, 1) how learning is conceptualized (as cognitive, behavioral, or social change), 2) how learners engage with models (by constructing them or by experimenting with pre-built ones), and 3) how learning is measured. Our central finding is that the three dimensions do not vary independently. Engagement mode strongly predicts both what a study treats as learning and how it measures it, so that the field divides into two largely non-overlapping traditions. Because study designs co-vary, findings do not accumulate across traditions leading to fragmentation often noted by the field. We close by identifying the combinations the literature seldom occupies, among them field evidence that model engagement improves real decisions, and social learning studies that connect shared understanding to behavior, and by arguing that a cumulative evidence base requires deliberately crossing these dimensions rather than continuing to travel within them.

[1] University College London
[2] Melbourne Business School
[3] University College London
[4] MIT
[5] Washington University in St. Louis

1. Introduction

Complex, dynamic systems are difficult to understand and to manage well (Sterman, 2010; Fischer, Degen & Funke, 2015). Decades of laboratory evidence show that people systematically misperceive feedback, are insensitive to delays, misjudge accumulation, and consequently generate oscillation and overshoot even in relatively simple managed systems (Arango Aramburo et al., 2012). System dynamics is, at its core, a response to this difficulty. It is a scientific approach "developed to facilitate our understanding and management of complex, dynamic systems" through modeling, simulation, and analysis (Davidsen & Spector, 2015). Its animating claim is that when people build or use formal models of feedback, accumulation, and delay, they revise the mental models they carry, and that these revised mental models support better reasoning and better decisions. The claim is foundational and one of th reasons we invest in group model building, in management flight simulators, and in classroom modeling. Yet, it is more often assumed than demonstrated and when it is tested, it is tested in strikingly different ways.

Following Sterman (2000), we take a mental model to be a person's internal representation of a system's causal structure, its variables, feedback loops, delays, and accumulations, together with the assumptions that person uses to reason about how the system behaves. Forrester (2016) makes the stakes of this definition plain: mental models "control nearly all social and economic activities" and are rich, reasonably reliable stores of information about a system's structure and policies, yet they have serious weaknesses including incompleteness, internal contradiction, and above all our inability to draw correct dynamic conclusions from them. Formal modeling and simulation are meant to compensate for our mental model shortcomings. Learning, in this review, is a change in that representation or in what it enables a person to do: to explain the system, to predict its behavior, to decide within it, or to agree with others about it. This deliberately broad definition is necessary because the studies we review do not share a narrower one; part of our task is to show how differently they operationalize the same word.

Prior reviews have tended to probe this territory deeply but narrowly. Kopainsky and Saldarriaga (2012), for instance, offer a careful synthesis of what single-player simulators reveal about individual understanding. A recurring lesson of that work is that understanding and performance are not straightforward proxies for one another. They write, "neither performance can be predicted by understanding nor can understanding be inferred from performance" (Kopainsky & Saldarriaga, 2012) and that the field still lacks methods that go beyond documenting novice–expert gaps to explain how learners acquire, transfer, and retain knowledge. Other reviews are similarly bounded by scope: Scott et al. (2016) synthesize the effectiveness of group model building specifically, finding it changes participants and supports consensus but has "not yet been translated into changing practice," while Groesser and Schaffernicht (2012) take stock of the mental-model construct itself. That focus buys analytical precision at the cost of breadth. Our aim is complementary but widens the scope from past reviews. We include both engagement modes rather than only simulators or model construction; cognitive, behavioral, and social conceptualizations of learning rather than understanding alone; a range of measurement approaches; and case studies alongside laboratory experiments. The contribution of the paper is twofold. First, we provide a structured map of a fragmented literature, organized around three dimensions that we define and then apply consistently to 32 studies. Second, and more analytically, we show that the dimensions are correlated. This co-variation is, we argue, the concrete form the field's "fragmentation" takes. That the field lacks a shared account of what is learned, or at what stage of proficiency, is itself a recognized problem by Schaffernicht and Groesser (2016), who argue that without an explicit framework, cumulative research on learning and teaching in system dynamics is hindered. So long as this fragmentation continues, accumulating evidence will be difficult.

2. Reviewing approach and coding framework

We assembled a corpus of 32 empirical studies published between 2010 and 2024, identified through database searches (Scopus, ProQuest, EBSCOhost) and searching of the System Dynamics Review, using term groups combining model based interventions (group model building, simulation, microworld, flight simulator, participatory modeling, causal mapping) with learning outcomes (learning, understanding, mental model, decision, consensus). We included empirical studies that examine learning from engagement with a system dynamics model. We excluded purely theoretical and literature review contributions, and studies in which a model serves only as a research instrument rather than as something a person learns from. Each study was coded along four dimensions using a codebook developed iteratively in NVivo. Three of these (conceptualization, engagement, and measurement) are the organizing axes of the review;

the fourth, study type, we treat as descriptive context. Note that codes are not mutually exclusive within a dimension, a study may conceptualize learning as both cognitive and behavioral change or combine two measurement approaches, for example.

2.1 The codebook

The codebook comprises four dimensions, each defined at the level of the dimension and then subdivided into codes. The dimensions are deliberately framed so that they interlock. An engagement mode activates particular learning processes, a conceptualization specifies which resulting change counts as learning, and a measurement approach makes that change observable. We define the four dimensions as follows, using the definitions applied during coding.

1. Engagement mode is the manner in which a learner interacts with a model, whether by constructing it or by experimenting with a pre-built one. This shapes what the learner does, which cognitive and social processes are activated, and therefore which forms of learning can occur and be observed. It divides into "active model construction" and "engagement with an existing model."

2. Conceptualization of learning is the theoretical stance defining what it means to learn from a system dynamics model -- which change (in an individual's cognition, in their behavior, or in a group's shared understanding) is treated as evidence that learning has occurred. It divides into cognitive, behavioral, and social conceptualizations.

3. Measurement approach is the method by which learning is made empirically observable. It is the operationalization that turns a claimed change in cognition, behavior, or shared understanding into evidence, and which therefore reflects the study's underlying conceptualization of learning. It divides into four measures: structural mental model change, dynamic reasoning or systems thinking, decision performance, and social learning or consensus.

4. Study type is the research design (experimental, case study, theoretical, or literature review) which we record as descriptive context rather than treat as an organizing axis. Table 1 gives the full code-level definitions within each dimension.

Dimension	Code	What it captures

Engagement	Active model construction	The learner helps build the model (group model building, community-based SD, client or classroom modeling). Learning is located in articulating and revising structure: externalizing tacit assumptions so they can be challenged, plus the cognitive work of reasoning about causality while building.
	Engagement with an existing model	The learner uses a pre-built model (microworlds, flight simulators, role-play). Learning is located in experimenting with behavior by making decisions, observing consequences quickly, and learning from the mismatch between expectation and outcome.
Conceptualization	Cognitive	Learning is a change in mental models and structural understanding: more accurate or more complete representations of causal structure, feedback, delays, and accumulation (sometimes deeper “conceptual change”).
	Behavioral	Learning is a change in decisions and performance: adapting decision rules to manage dynamic complexity more effectively.
	Social	Learning is a change in collective understanding: alignment of causal interpretations, consensus, and shared sensemaking across a group.
Measurement	Structural measures of mental-model change	Learners externalize their understanding as causal loop diagrams or concept maps, which are converted to adjacency matrices and compared against an expert reference model, scoring variables, link polarity, feedback loops, and delays.
	Measures of dynamic reasoning / ST	Behavioral tasks testing whether people can infer system behavior — e.g., stock-and-flow prediction tasks and think-aloud protocols.
	Decision-performance metrics	Performance in dynamic tasks and simulators: target achievement, system stability, and robustness across rounds.

	Measures of social learning / consensus	Map-similarity indices across participants and discourse analysis of how agreement forms.
Study type	Experimental / Case study / Theoretical / Literature review	The research design. Literature reviews supply framing and background but are not part of the analysis proper; study type is recorded as descriptive context rather than an organizing axis.

Table 1. The four-dimension coding framework applied to the 32 studies.

The corpus is unevenly distributed across these codes (Table 2), and the pattern is itself informative. Learning is conceptualized cognitively in the large majority of studies and behaviorally or socially in fewer; it is studied experimentally far more often than in the field; and the measure best matched to the social conceptualization (social learning and consensus) is the rarest of all.

Code	n	%	Dimension
Cognitive	27	84%	Conceptualization
Experimental	23	72%	Study type
Engagement with existing model	19	59%	Engagement
Active model construction	16	50%	Engagement
Dynamic reasoning / ST	13	41%	Measurement
Structural mental-model change	13	41%	Measurement
Behavioral	12	38%	Conceptualization
Social	9	28%	Conceptualization
Case study	9	28%	Study type
Decision performance	8	25%	Measurement
Social learning / consensus	5	16%	Measurement

Table 2. Code frequencies across the 32 studies (a study may carry several codes per dimension).

2.2 Coding, validation, and scope

Two features of the coding process influence the credibility of the patterns reported below. The first is scope. An initial pool of 54 candidate studies was reduced to the 32 analyzed in this manuscript. We aside conceptual pieces in which a model serves only as a research instrument rather than as something a person learns from and excluded papers identified in the search but not obtainable in full text rather than coding them from abstracts. The codebook was developed and applied iteratively by a single coder, and every study was coded a second time from full text and then cross-checked against an earlier round of coding, with each discrepancy resolved by returning to the source. We make no claim for formal inter-rater reliability. Coding reflects the analyst's judgment against the explicit definitions in Table 1, and individual cell assignments should be read as interpretive rather than definitive. Because codes are not mutually exclusive, a study that genuinely spans categories is recorded in each, which relieves the pressure to force ambiguous cases into a single box and makes the category-spanning itself analyzable. The corpus reflects English language work in the searched databases and the field's own publication tendencies.

3. How the literature conceptualizes learning

The first and most consequential source of variation is what a study takes learning to be. We distinguish three conceptualizations (cognitive, behavioral, and social) that correspond to three kinds of change a study may treat as evidence that learning has occurred: 1) change in an individual's understanding, 2) change in an individual's decisions, and 3) change in a group's shared understanding. Most studies foreground one, though a substantial minority combine two.

3.1 Cognitive: mental models and structural understanding

The dominant conceptualization treats learning as improved dynamic reasoning and/or system thinking skills. Present in 27 of the 32 studies (84%), it is the lens of the stock-and-flow literature, in which people are asked to infer how accumulations behave under given flows (Sterman, 2010; Baghaei Lakeh & Ghaffarzadegan, 2015, 2016; Fischer, Degen & Funke, 2015; Aşık & Doğança Küçük, 2021). The cognitive code underlies studies that elicit a learner's causal map and compare it to an expert's, too (Plate, 2010; Elsawah et al., 2017; Gary & Wood, 2016). It is also the strand that frames model-based learning explicitly as conceptual change, i.e. the restructuring of prior, often flawed, intuitions about dynamics (Lee, 2010; Jensen, 2014; Van Borkulo et al., 2012; Wijnen et al., 2015). The wager of this lens is that structural improvement pays off: Gary and Wood (2016) report that more accurate mental models yield better decisions and higher performance, adding the important qualification that accurate mental models of key dynamics (rather than the whole system) are sufficient. Cognition, including mental models, are therefore "invaluable for analyzing how individuals and groups approach complex problems" (Doyle & Ford, 1998, as cited in Haque et al., 2023). A recurring and sobering finding must be noted, though: learners' mental models remain far less accurate than experts', even after instruction (Elsawah et al., 2017; McCardle-Keurentjes et al., 2018), and those gains can be fragile.

3.2 Behavioral: decisions, strategies, and performance

A second conceptualization treats learning as improved performance, meaning better decisions in a dynamic task, indexed by how close a learner comes to a target, how stable the managed system remains, or how robust performance is across rounds (Yang et al., 2016; Brauch, 2022; Borštnar et al., 2011; Kopainsky & Sawicka, 2011; Öllinger et al., 2015). Performance is itself systematically shaped by dynamic complexity. As the complexity of the managed environment rises, results fall away from optimal benchmarks, and the best performing teams tend to be those that better grasp the system's structure (Arango Aramburo et al., 2012). This lens is most at home in simulator studies, and it carries a well-known interpretive hazard that several of its own practitioners raise: improved performance does not indicate improved understanding. Gary and Wood (2016) show that partial knowledge of a few key feedbacks can be enough to decide well, while gains elsewhere may reflect procedural familiarity rather than structural insight. Learning defined behaviorally and learning defined cognitively are therefore not interchangeable, even when they are measured on the same task.

3.3 Social: consensus and shared sensemaking

A third conceptualization locates learning not in the individual but in the group -- in the alignment of participants' causal interpretations, the emergence of consensus, and the construction of shared meaning. It is the lens of the group model building and participatory literatures (Scott et al., 2013; Scholz et al., 2015; Herrera et al., 2016; Thompson et al., 2016; Rouwette et al., 2011; Rooney-Varga et al., 2021), and of conceptual work that treats models and diagrams as boundary objects around which shared understanding is negotiated (Black, 2013, 2024; Richardson & Andersen, 2010; Midgley et al., 2013). On this view the aim of a model, whether a co-created causal diagram or a fixed, expertly calibrated simulator such as En-ROADS or the Re-Think Health microworld, is less to transfer a correct structure than to foster "intelligent action among the stakeholders facing the problem" (Black, 2024). Here, Black reframes the guiding question as one of audience: useful to whom, and who must act on what they learn. The evidence here is suggestive. Group model building often improves communication and reported agreement but it is also the least secured. Rouwette et al. (2011) caution that there is "no clear evidence for the effectiveness of group model building" and that participants often do not even recognize that their mental models have changed. Herrera et al. (2016) add that because most evidence comes from single cases evaluated with different criteria, the field needs controlled comparison before extrapolating. Consensus itself is an ambiguous signal, since convergence may reflect genuine shared understanding or merely conformity and compromise.

4. How learners engage with models

The second dimension concerns what the learner actually does. We distinguish two modes, understanding them less as a dichotomy than as the ends of a continuum with hybrid designs in between. The distinction matters because, as we show in Section 6, it is the strongest predictor of how a study defines and measures learning.

4.1 Active model construction

In construction approaches, the learner helps build the model. This includes group and community based system dynamics with stakeholders (Scott et al., 2013; Scholz et al., 2015; Herrera et al., 2016; Rouwette et al., 2011; Thompson et al., 2016; Moallemi et al., 2021), classroom and curricular modeling in which students externalize and revise causal structure (Lee, 2010; Jensen, 2014; Van Borkulo et al., 2012; Wijnen et al., 2015; Dhawan et al., 2011; Sterman, 2010), and methodological work on how such externalized models can be compared and scored (Haque et al., 2023). The theorized mechanism is that articulating structure surfaces tacit assumptions so they can be examined and revised -- learning through the act of representation. Consistent with that mechanism, construction studies overwhelmingly define learning cognitively or socially rather than behaviorally.

4.2 Engagement with existing models

In the second mode the learner experiments with a model someone else has built -- a microworld, a management flight simulator, an interactive learning environment, or a role-play built on a policy model (Yang et al., 2016; Brauch, 2022; Gary & Wood, 2011, 2016; Öllinger et al., 2015; Kopainsky & Sawicka, 2011; Bußwolder, 2015; Elsawah et al., 2017; Rooney-Varga et al., 2018, 2021; Sterman et al., 2015). Here learning is theorized to come from rapid structure-behavior feedback. The learner makes a decision, sees its consequences quickly, and updates from the mismatch between expectation and outcome. A consistent qualification across this literature is that experimentation alone is not sufficient, what learners take from a simulator depends heavily on interface and scaffolding design, including progressive complexity, structural transparency, and guided debriefing (Kopainsky & Sawicka, 2011; Elsawah et al., 2017; Pavlov et al., 2015).

A few studies deliberately combine the modes by having learners both build and run models or comparing construction against pre-built alternatives within one design (McCardle-Keurentjes et al., 2018; Mulder et al., 2015; Wijnen et al., 2015). These are the exception. Only three of the 32 studies engage both modes; the overwhelming majority commit to one. Direct, controlled comparison of construction versus existing-model engagement, the design that could tell us which mode produces which kind of learning, is therefore almost absent.

5. How learning is measured

The third dimension concerns operationalization: how a claimed change is made observable. Four approaches recur, and each is closely tied to a particular conceptualization of learning.

5.1 Structural measures of mental model change

The most direct cognitive measure, used in 13 of the 32 studies (41%), asks learners to externalize their understanding as a causal loop diagram or concept map, converts it to an adjacency matrix, and scores it against an expert reference on variables, link polarity, feedback loops, and delays (Plate, 2010; Gary & Wood, 2016; Bußwolder, 2015; Elsawah et al., 2017; McCardle-Keurentjes et al., 2018; Öllinger et al., 2015; Scholz et al., 2015; Rouwette et al., 2011). It is the measure most faithful to the field’s own theory of learning, but it is not without problems. Haque et al. (2023) show that different network metrics applied to the same maps can yield inconsistent verdicts, and the coding of concepts is sensitive to the level of abstraction chosen.

5.2 Measures of dynamic reasoning and systems thinking

A second family of measures tests whether people can reason about system behavior directly, through stock-and-flow prediction tasks and think-aloud protocols (Sterman, 2010; Baghaei Lakeh & Ghaffarzadegan, 2015, 2016; Fischer, Degen & Funke, 2015; Aşık & Doğança Küçük, 2021; Dhawan et al., 2011; Richert et al., 2017; Kopainsky & Sawicka, 2011). These tasks often dispense with a model altogether, asking the learner to reason on paper or aloud to a researcher.

5.3 Decision performance measures

The behavioral conceptualization is operationalized through performance in a dynamic task including target achievement, system stability, or robustness across rounds in a simulator (8 of 32, 25%: Yang et al., 2016; Brauch & Größler, 2022; Borštnar et al., 2011; Gary & Wood, 2016; Öllinger et al., 2015; Kopainsky & Sawicka, 2011; McCardle-Keurentjes et al., 2018; Bußwolder, 2015). Performance measures are unambiguous and comparable, which is their appeal; but, as noted, they cannot on their own distinguish understanding from procedural skill (e.g. learning to play the game), and, as we show below, they are collected almost exclusively in the laboratory.

5.4 Measures of social learning and consensus

Finally, the social conceptualization is operationalized through map similarity indices across participants and through discourse analysis of how agreement forms (Scholz et al., 2015; Scott et al., 2013; Herrera et al., 2016; Thompson et al., 2016; Rooney-Varga et al., 2021). This is the rarest measurement approach in the corpus (five studies coming in at 16%) and it is never paired with a decision-performance measure and only once with a behavioral conceptualization (Rooney-Varga et al., 2021), so that shared understanding is seldom connected to whether groups subsequently decide or act differently after engaging with a model.

6. Where do we stand?

Reviewing the three dimensions separately, as we have just done, understates the most important pattern in the corpus, which appears only when they are crossed. The dimensions do not vary freely. Engagement mode strongly predicts both how a study conceptualizes learning and how it measures it, so that the field sorts into two bundles.

The first alignment is between engagement and conceptualization (Table 3). Existing model studies lean behavioral (11 of 19) and rarely social (4 of 19); construction studies almost never define learning behaviorally (2 of 16) and carry the social lens somewhat more often (5 of 16). The cognitive conceptualization sits on both sides as the connective tissue of the field, the one thing nearly every study shares.

	Behavioral	Cognitive	Social
Active construction	2	14	5
Existing model	11	16	4

Table 3. Engagement mode × conceptualization of learning (counts of studies).

The second alignment carries the pattern into measurement (Table 4). Existing model studies reach for decision performance metrics (8 of 19); construction studies spread across structural, consensus, and reasoning measures and almost never use performance (1 of 16). And measurement follows conceptualization almost deterministically: behavioral learning is measured as decision performance, cognitive learning as dynamic reasoning or structural mental model change, social learning as consensus -- the last of which never appears outside the social lens.

	Decision perf.	Dynamic reas.	Social learn.	Struct. MM
Active construction	1	5	4	8
Existing model	8	8	1	8

Table 4. Engagement mode × measurement approach (counts of studies).

Put together, the two alignments describe two self-contained research traditions, summarized in Table 5. In the first, a learner engages a pre-built model, learning is defined as better decisions, it is measured as task performance, and the study is a laboratory experiment. In the second, learners build a model together, learning is defined as shared or cognitive understanding, it is measured as consensus or map structure, and the study is a field case. A third, cognitive strand (the stock-and-flow and systems-thinking work) cuts across the engagement distinction, appearing under both modes, and supplies much of the field's connective cognitive vocabulary without belonging cleanly to either tradition.

	Tradition A: existing model	Tradition B: construction
Engagement	Experiment with a pre-built model (microworlds, flight simulators, role-play)	Build a model together (group model building, participatory, classroom)
Learning is…	Better decisions (behavioral), sometimes better understanding	Shared or restructured understanding (social, cognitive)
Measured as…	Decision performance in a dynamic task	Consensus / map similarity; structural mental-model change
Setting	Laboratory experiment	Field case study
Typical studies	Yang 2016; Brauch & Größler 2022; Gary & Wood 2016; Öllinger 2015	Scott 2013; Scholz 2015; Moallemi 2021; Thompson 2016

Table 5. Two aligned research traditions, plus a cross-cutting cognitive strand.

This is, we suggest, the concrete form the field's fragmentation takes. It is not merely that studies are diverse; it is that their design choices are correlated, so that each tradition tests its own version of the premise with its own instruments in its own setting. Findings therefore do not accumulate across the divide. A performance result from a simulator experiment and a consensus result from a group model building case are not competing evidence about one question but parallel evidence about two. Recognizing this does not diminish either tradition (both have produced careful, informative work), but it explains why a field with a shared premise can nonetheless struggle to build a shared evidence base.

We state the pattern as four propositions.

- Proposition 1. Engagement mode predicts the conceptualization of learning: studies that have learners experiment with pre-built models tend to define learning behaviorally, whereas studies in which learners construct models tend to define it cognitively or socially.
- Proposition 2. Conceptualization predicts measurement: behavioral learning is operationalized as decision performance, cognitive learning as dynamic-reasoning or structural measures, and social learning as consensus.
- Proposition 3. Because these alignments compound, the field sorts into two research traditions plus a cross-cutting cognitive strand. It is this bundling of design choices that makes findings non-cumulative across traditions.
- Proposition 4. It follows that the most informative studies are those that break the alignment by pairing an engagement mode with an atypical conceptualization, measure, or setting.

7. Strengths and limitations of the evidence base

Two observations from within the field frame the assessment that follows. McCardle-Keurentjes et al. (2018) point to an asymmetry in six decades of experimental work: how people come to understand system behavior has been studied extensively, whereas how they come to understand system structure "has been neglected," and the core assumption that decision aids such as causal loop diagrams and group model building actually build richer mental models "remain[s]" largely untested. Davidsen and Spector (2015), reflecting on more than a decade of progress, observe that although simulation technologies have grown enormously in power, "no corresponding growth in terms of improving learning, decision making and performance has occurred." Capability and demonstrated learning, in other words, are not the same thing.

The literature has real strengths. It is methodologically diverse, spanning controlled experiments, field cases, and conceptual work; it is applied, tied to genuine problems in education, resource management, and climate policy (Rooney-Varga et al., 2018, 2021; Sterman et al., 2015; Moallemi et al., 2021); and within each tradition its instruments have matured, from validated stock-flow tasks to increasingly sophisticated map-comparison methods.

Its limitations follow from the alignment we have described. Because “learning” is conceptualized inconsistently and rarely situated in an explicit theoretical framework, studies evaluate different phenomena under one label. Because performance and understanding are usually measured separately, the field cannot easily tell procedural gains from structural insight (Gary & Wood, 2016; Bußwolder, 2015; Öllinger et al., 2015; McCardle-Keurentjes et al., 2018). And because almost all measurement is short-horizon, the durability and transfer of learning are largely untested; whether brief interventions produce enduring change in what are, on dual-process accounts, relatively stable deliberative structures remains an open question. We note that our coding cannot itself adjudicate durability, since every measurement code names an outcome type rather than a time horizon; establishing it would require a temporal coding pass over the same corpus.

8. Where should we go?

If the dimensions co-vary, then the most valuable studies are those that deliberately break the alignment, occupying combinations the field has left nearly empty. Four such gaps stand out.

First, decision performance is measured only in the laboratory. All eight performance studies are experiments; the code never co-occurs with field or case work. The field therefore has no field-context evidence that engaging with a model improves real-world decisions. Performance gains exist only where a controlled task defines them. Studies that carry a performance measure into an applied setting would test the premise where it ultimately matters. Consistent with this, a scoping review of system dynamics in environmental health policy finds applied uptake "currently limited," with barriers that are not specific to that domain but generic to the method (Currie et al., 2018).

Second, social learning is asserted more than measured and is barely connected to behavior. Nine studies adopt the social lens but only five measure consensus or shared understanding. A social learning measure never co-occurs with a decision performance one, and it is paired with a behavioral conceptualization in just a single study. Whether shared understanding actually changes how groups decide or act is, at present, all but unstudied -- a natural priority given how much of the field’s practice is participatory.

Third, the two engagement modes are almost never compared head-to-head. With only three of 32 studies engaging both modes, the field’s repeated claim that construction and simulation produce different kinds of learning rests largely on cross-study inference rather than within-study comparison. Controlled comparisons (same domain, same measures, construction versus pre-built engagement) would let the field attribute outcomes to mode rather than to the many things that differ between the two traditions.

Fourth, durability and transfer remain open. Most studies measure minutes or hours after intervention; longitudinal designs that revisit learners weeks or months later, and transfer designs that test structurally analogous problems, would establish whether model based learning endures and generalizes or merely activates knowledge briefly. Kopainsky and Saldarriaga (2012) name the relevant phenomena precisely -- knowledge acquisition, transfer, and retention -- and argue that the field has adequate methods for the first but few for the latter two. Herrera et al. (2022), reflecting on the long-lasting organizational effects of group model building, likewise note that how modelling insights durably change behavior and performance "remain[s] largely puzzling."

In practical terms, these are largely designs within reach. It may be a matter of recombining instruments the field already has rather than inventing new ones. Field tests of performance can partner with organizations that already use models in practice, using matched or stepped-wedge comparison groups and objective decision outcomes in place of simulator scores. Studies that connect social learning to behavior can pair existing consensus and map-similarity measures with downstream indicators like the decisions a group subsequently makes, or the changes it implements. Head-to-head comparisons of engagement mode can randomize participants to construction versus pre-built conditions while holding the domain, the model content, and the elicitation instrument constant, so that outcome differences are attributable to mode rather than to context. And durability can be probed with follow-up measurement at intervals of weeks or months, combined with transfer tasks on structurally analogous problems.

Underlying all four is a single recommendation. The field does not primarily need more studies within its established packages; it needs studies that cross the dimensions, that pair a conceptualization with an unusual measure, or an engagement mode with an unusual setting. And it needs the explicit theoretical frameworks that would justify those pairings by linking specific learning constructs to specific engagement mechanisms. Only then can goals, methods, and measures be aligned deliberately rather than by convention.

9. Conclusion

System dynamics assumes that engaging with models helps people learn about complex systems. Reviewing 32 studies that examine this assumption, we have organized the evidence along three dimensions: 1) how learning is conceptualized, 2) how learners engage with models, and 3) how learning is measured. The picture that emerges is not one of scattered diversity but of structured division: engagement mode predicts conceptualization and measurement, sorting the field into two aligned traditions bridged by a cognitive strand. That alignment is why a field with a common premise accumulates evidence so unevenly. The path to a more cumulative science of learning from models runs, we have argued, through the combinations the field has yet to occupy -- field tests of performance, social learning linked to behavior, direct comparison of engagement modes, and studies that follow learners over time. By deliberately crossing the dimensions along which the literature is currently aligned, the field can begin to build evidence that travels.

References

Arango Aramburo, S., Castañeda, J.A. and Olaya, Y. (2012) ‘Laboratory experiments in the system dynamics field’, System Dynamics Review, 28(1), pp. 94–106. doi: 10.1002/sdr.472.

Aşık, G. and Doğança Küçük, Z. (2021) ‘Metacognition in action as a possible explanation for stock-flow failure’, System Dynamics Review, 37(4), pp. 253–282. doi: 10.1002/sdr.1692.

Baghaei Lakeh, A. and Ghaffarzadegan, N. (2015) ‘Does analytical thinking improve understanding of accumulation?’, System Dynamics Review, 31(1–2), pp. 46–65. doi: 10.1002/sdr.1528.

Baghaei Lakeh, A. and Ghaffarzadegan, N. (2016) ‘The dual-process theory and understanding of stocks and flows’, System Dynamics Review, 32(3–4), pp. 309–331. doi: 10.1002/sdr.1566.

Barnabè, F. and Davidsen, P.I. (2020) ‘Exploring the potentials of behavioral system dynamics: insights from the field’, Journal of Modelling in Management, 15(1), pp. 339–364. doi: 10.1108/JM2-03-2019-0081.

Bastani, R. (2016) 'Enhancing children's understanding of complex systems: a study of the design considerations of an effective learning environment', in Proceedings of EdMedia 2016. Waynesville, NC: Association for the Advancement of Computing in Education (AACE), pp. 1348–1353.

Black, L.J. (2013) 'When visuals are boundary objects in system dynamics work', System Dynamics Review, 29(2), pp. 70–86. doi: 10.1002/sdr.1496.

Black, L.J. (2024) 'Reflecting on Zagonel's dichotomy of microworlds and boundary objects', System Dynamics Review, 40(4), e1783. doi: 10.1002/sdr.1783.

Borštnar, K.M., Kljajić, M., Škraba, A., Kofjač, D. and Rajkovič, V. (2011) 'The relevance of facilitation in group decision making supported by a simulation model', System Dynamics Review, 27(3), pp. 270–293. doi: 10.1002/sdr.460.

Brauch, M. and Größler, A. (2022) 'Holistic versus analytic thinking orientation and its relationship to the bullwhip effect', System Dynamics Review, 38(2), pp. 121–134. doi: 10.1002/sdr.1702.

Bußwolder, P. (2015) 'The effect of a structured method on mental model accuracy and performance in a complex task', Systems, 3(4), pp. 264–286. doi: 10.3390/systems3040264.

Currie, D.J., Smith, C. and Jagals, P. (2018) 'The application of system dynamics modelling to environmental health decision-making and policy – a scoping review', BMC Public Health, 18, 402. doi: 10.1186/s12889-018-5318-8.

Davidsen, P.I. and Spector, J.M. (2015) 'Critical reflections on system dynamics and simulation/gaming', Simulation & Gaming, 46(3–4), pp. 430–444. doi: 10.1177/1046878115596526.

Dhawan, R., O'Connor, M. and Borman, M. (2011) 'The effect of qualitative and quantitative system dynamics training: an experimental investigation', System Dynamics Review, 27(3), pp. 313–327. doi: 10.1002/sdr.455.

Doyle, J.K. and Ford, D.N. (1998) 'Mental models concepts for system dynamics research', System Dynamics Review, 14(1), pp. 3–29. doi: 10.1002/(SICI)1099-1727(199821)14:1<3::AID-SDR140>3.0.CO;2-K.

Elsawah, S., McLucas, A. and Mazanov, J. (2017) 'An empirical investigation into the learning effects of management flight simulators: a mental models approach', European Journal of Operational Research, 259(1), pp. 262–272. doi: 10.1016/j.ejor.2016.10.011.

Fischer, H., Degen, C. and Funke, J. (2015) 'Improving stock-flow reasoning with verbal formats', Simulation & Gaming, 46(3–4), pp. 255–269. doi: 10.1177/1046878114565058.

Forrester, J.W. (2016) 'Learning through system dynamics as preparation for the 21st century', System Dynamics Review, 32(3–4), pp. 187–203. doi: 10.1002/sdr.1571.

Gary, M.S. and Wood, R.E. (2016) 'Unpacking mental models through laboratory experiments', System Dynamics Review, 32(2), pp. 99–127. doi: 10.1002/sdr.1560.

Groesser, S.N. and Schaffernicht, M. (2012) 'Mental models of dynamic systems: taking stock and looking ahead', System Dynamics Review, 28(1), pp. 46–68. doi: 10.1002/sdr.476.

Haque, S., Mahmoudi, H., Ghaffarzadegan, N. and Triantis, K. (2023) 'Mental models, cognitive maps, and the challenge of quantitative analysis of their network representations', System Dynamics Review, 39(2), pp. 152–170. doi: 10.1002/sdr.1729.

Herrera, H.J., McCardle-Keurentjes, M.H.F. and Videira, N. (2016) 'Evaluating facilitated modelling processes and outcomes: an experiment comparing a single and a multimethod approach in group model building', Group Decision and Negotiation, 25(6), pp. 1277–1318. doi: 10.1007/s10726-016-9480-z.

Herrera, H., Videira, N., Korzilius, H.P.L.M., Cordova-Pozo, K.L. and McCardle-Keurentjes, M.H.F. (2022) 'Reflecting on factors influencing long-lasting organisational effects of group model-building interventions', System Dynamics Review, 38(2), pp. 190–209. doi: 10.1002/sdr.1705.

Jensen, E. (2014) 'Does teaching students how to explicitly model the causal structure of systems improve their understanding?' European Journal of Engineering Education, 39:4, 391-411, DOI: 10.1080/03043797.2014.881320

Kopainsky, B. and Saldarriaga, M. (2012) 'Assessing understanding and learning about dynamic systems', in Proceedings of the 30th International Conference of the System Dynamics Society. St. Gallen, Switzerland: System Dynamics Society.

Kopainsky, B. and Sawicka, A. (2011) 'Simulator-supported descriptions of complex dynamic problems: experimental results on task performance and system understanding', System Dynamics Review, 27(2), pp. 142–172. doi: 10.1002/sdr.445.

Lee, C.B. (2010) ‘The interactions between problem solving and conceptual change: system dynamic modelling as a platform for learning’, Computers & Education, 55(3), pp. 1145–1158. doi: 10.1016/j.compedu.2010.05.012.

McCardle-Keurentjes, M.H.F., Rouwette, E.A.J.A., Vennix, J.A.M. and Jacobs, E. (2018) ‘Potential benefits of model use in group model building: insights from an experimental investigation’, System Dynamics Review, 34(1–2), pp. 354–384. doi: 10.1002/sdr.1603.

Midgley, G., Cavana, R.Y., Brocklesby, J., Foote, J.L., Wood, D.R.R. and Ahuriri-Driscoll, A. (2013) ‘Towards a new framework for evaluating systemic problem structuring methods’, European Journal of Operational Research, 229(1), pp. 143–154. doi: 10.1016/j.ejor.2013.01.047.

Moallemi, E.A. et al. (2021) ‘Evaluating participatory modeling methods for co-creating pathways to sustainability’, Earth’s Future, 9(3), e2020EF001843. doi: 10.1029/2020EF001843.

Mulder, Y.G., Lazonder, A.W. and de Jong, T. (2015) ‘Simulation-based inquiry learning and computer modeling: pitfalls and potentials’, Simulation & Gaming, 46(3–4), pp. 322–347. doi: 10.1177/1046878115577159.

Öllinger, M., Hammon, S., von Grundherr, M. and Funke, J. (2015) ‘Does visualization enhance complex problem solving? The effect of causal mapping on performance in the computer-based microworld Tailorshop’, Educational Technology Research and Development, 63(4), pp. 621–637. doi: 10.1007/s11423-015-9393-6.

Pavlov, O.V., Saeed, K. and Robinson, L.W. (2015) ‘Improving instructional simulation with structural debriefing’, Simulation & Gaming, 46(3–4), pp. 383–403. doi: 10.1177/1046878114567980.

Plate, R. (2010) ‘Assessing individuals’ understanding of nonlinear causal structures in complex systems’, System Dynamics Review, 26(1), pp. 19–33. doi: 10.1002/sdr.432.

Richardson, G.P. and Andersen, D.F. (2010) ‘Systems thinking, mapping, and modeling in group decision and negotiation’, in Eden, C. and Kilgour, D.M. (eds) Handbook of Group Decision and Negotiation. Dordrecht: Springer, pp. 313–324. doi: 10.1007/978-90-481-9097-3_19.

Richert, C., Boschetti, F., Walker, I., Price, J. and Grigg, N. (2017) ‘Testing the consistency between goals and policies for sustainable development: mental models of how the world works today are inconsistent with mental models of how the world will work in the future', Sustainability Science, 12(1), pp. 45–64. doi: 10.1007/s11625-016-0384-2.

Rooney-Varga, J.N., Sterman, J.D., Fracassi, E., Franck, T., Kapmeier, F., Kurker, V., Johnston, E., Jones, A.P. and Rath, K. (2018) 'Combining role-play with interactive simulation to motivate informed climate action: evidence from the World Climate simulation', PLOS ONE, 13(8), e0202877. doi: 10.1371/journal.pone.0202877.

Rooney-Varga, J.N., Hensel, M., McCarthy, C., McNeal, K.S., Norfles, N., Rath, K., Schnell, A.H. and Sterman, J.D. (2021) 'Building consensus for ambitious climate action through the World Climate simulation', Earth's Future, 9(12), e2021EF002283. doi: 10.1029/2021EF002283.

Rouwette, E.A.J.A., Korzilius, H., Vennix, J.A.M. and Jacobs, E. (2011) 'Modeling as persuasion: the impact of group model building on attitudes and behavior', System Dynamics Review, 27(1), pp. 1–21. doi: 10.1002/sdr.441.

Schaffernicht, M.F.G. and Groesser, S.N. (2016) 'A competence development framework for learning and teaching system dynamics', System Dynamics Review, 32(1), pp. 52–81. doi: 10.1002/sdr.1550.

Scholz, G., Austermann, M., Kaldrack, K. and Pahl-Wostl, C. (2015) 'Evaluating group model building exercises: a method for comparing externalized mental models and group models', System Dynamics Review, 31(1–2), pp. 28–45. doi: 10.1002/sdr.1538.

Scott, R.J., Cavana, R.Y. and Cameron, D. (2013) 'Evaluating immediate and long-term impacts of qualitative group model building workshops on participants' mental models', System Dynamics Review, 29(4), pp. 216–236. doi: 10.1002/sdr.1505.

Scott, R.J., Cavana, R.Y. and Cameron, D. (2016) 'Recent evidence on the effectiveness of group model building', European Journal of Operational Research, 249(3), pp. 908–918. doi: 10.1016/j.ejor.2015.06.078.

Sterman, J.D. (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. Boston, MA: Irwin/McGraw-Hill.

Sterman, J.D. (2010) 'Does formal system dynamics training improve people's understanding of accumulation?', System Dynamics Review, 26(4), pp. 316–334. doi: 10.1002/sdr.447.

Sterman, J., Franck, T., Fiddaman, T., Jones, A., McCauley, S., Rice, P., Sawin, E., Siegel, L. and Rooney-Varga, J.N. (2015) 'World Climate: a role-play simulation of climate negotiations', Simulation & Gaming, 46(3–4), pp. 348–382. doi: 10.1177/1046878113514935.

Thompson, J.P., Howick, S. and Belton, V. (2016) 'Critical learning incidents in system dynamics modelling engagements', European Journal of Operational Research, 249(3), pp. 945–958. doi: 10.1016/j.ejor.2015.09.048.

van Borkulo, S.P., van Joolingen, W.R., Savelsbergh, E.R. and de Jong, T. (2012) 'What can be learned from computer modeling? Comparing expository and modeling approaches to teaching dynamic systems behavior', Journal of Science Education and Technology, 21(2), pp. 267–275. doi: 10.1007/s10956-011-9314-3.

Wijnen, F.M., Mulder, Y.G., Alessi, S.M. and Bollen, L. (2015) 'The potential of learning from erroneous models: comparing three types of model instruction', System Dynamics Review, 31(4), pp. 250–270. doi: 10.1002/sdr.1546.

Yang, M.M., Jiang, H. and Gary, M.S. (2016) 'Challenging learning goals improve performance in dynamically complex microworld simulations', System Dynamics Review, 32(3–4), pp. 204–232. doi: 10.1002/sdr.1559.