Reading the Map: A Dual-Lens Protocol for Assessing Systems Thinking Artifacts from Brief Workshops in Engineering Entrepreneurial Settings

Amin Azad[1,*] and Emily Moore[1]

[1]Institute for Studies in Transdisciplinary Engineering Education and Practice (ISTEP), University of Toronto, Toronto, ON M5S 1A1, Canada
[*]Correspondence: amin.azadarmaki@mail.utoronto.ca

Abstract

Introduction. Assessing what learners actually take away from a brief systems thinking workshop is harder than it looks. Most published evaluations rely on post session self-report, which is quick to administer but vulnerable to response shift, social desirability, and the simple fact that participants have just spent four hours invested in the very intervention they are being asked to evaluate. The Iceberg models and causal loop diagrams that learners produce during these workshops carry visible traces of how they reason, such as which patterns they name, which actors they include, and which feedback loops they specify, but these artifacts are rarely used as evidence on their own terms. The gap matters particularly in entrepreneurial settings, where teams reason about live ventures rather than instructor-controlled cases and where self-report is at its most fragile.

Approach. We present a dual lens protocol for assessing the systems thinking artifacts that brief workshops generate. Lens1 defines Richmond's systems thinking skills (dynamic, system as cause, closed loop, forest, and operational thinking) as features that can be recognized on the artifact itself without knowing the venture's technical domain. Lens2 defines opportunity identification signals: whether teams converted event level symptoms into structural drivers, expanded boundaries to include intermediaries and gatekeepers, identified leverage points within feedback structure, and surfaced feasibility or adoption constraints. We demonstrate the protocol on twelve artifacts from a four hour workshop with seven early stage engineering venture teams.

Results. The dual lens separates artifacts that genuinely demonstrate systems thinking from those that stop at surface description. Applied to the demonstration set, it surfaces concrete evidence of dynamic and closed loop reasoning, identifies which teams reached the leverage hypothesis stage, and exposes the kind of indicator level variation that aggregate self-report measures compress into a single number. Once the indicators are defined, the protocol can be applied by a single rater in under twenty minutes per artifact, with no venture domain expertise required.

Discussion. The protocol offers a behavioral, replicable instrument that complements rather than replaces survey based measures, and that addresses the assessment fragmentation flagged in recent reviews of systems thinking education. It is most useful in settings where authentic content is non-negotiable and self-report alone is insufficient. Three things still need to happen for broader adoption: independent rater reliability testing, an active control comparison against artifacts produced under non-systems thinking interventions, and longitudinal application across repeated workshop sittings.

Use of AI. AI tools were used for transcription assistance during a parallel evaluation and for minor language editing throughout. The authors designed the protocol, scored the artifacts, and wrote the paper, and take responsibility for the final content.

KEYWORDS: *systems thinking assessment; artifact based assessment; iceberg model; causal loop diagrams; engineering entrepreneurship education; methodology; brief interventions*

1 Introduction

Brief systems thinking workshops have been widely used in engineering and entrepreneurship education to introduce learners to interdependence, feedback, and structural reasoning within compact, time bounded formats (Senge, 1990; Sterman, 2000). A recurring challenge for educators and researchers delivering these interventions is how to assess what learners actually take away from them. The dominant evaluation strategy in published studies has been self-report on perceived understanding or perceived relevance, often collected immediately post workshop (Dugan et al., 2022). Self-report instruments are quick to deploy, but are systematically vulnerable to response shift bias and social desirability, and they cannot directly observe the reasoning moves the workshop is supposed to cultivate. The participant-produced artifacts that brief workshops typically generate, including Iceberg models and causal loop diagrams (CLDs), are rarely treated as evidence in their own right. These artifacts carry observable traces of the reasoning moves that the workshop targets: how learners separate event level symptoms from recurring patterns, how they push explanation downward into structural drivers, how they specify feedback structure, and how they connect any of this to identifiable intervention pathways (Richmond, 1993; Sterman, 2000). A recent systematic review of systems thinking assessments in engineering reports that artifact based instruments remain underdeveloped relative to survey and scenario based approaches and that the field would benefit from instruments that examine behavioral products in addition to perceived gains (Dugan et al., 2022; Grohs et al., 2018).

This gap is acute in entrepreneurial settings. Early stage venture teams reason under uncertainty about real ventures rather than instructor selected cases, which means any assessment must be interpretable across heterogeneous artifact content without requiring the rater to be an expert in the venture's technical domain. Entrepreneurship scholarship frames opportunity identification and development as an iterative process in which teams interpret cues, choose boundaries, articulate assumptions, and refine opportunity beliefs through action and learning (Ardichvili et al., 2003; Clausen, 2020). It also means that self-report is at its most vulnerable, because participants have strong motivational reasons to overstate the value of any intervention with which they have just spent four hours engaging (Howard & Dailey, 1979). If a workshop is intended to influence that reasoning, an assessment instrument should be able to observe its products directly.

This paper contributes a dual lens assessment protocol for systems thinking artifacts produced in brief workshops with early stage engineering venture teams, and demonstrates it on artifacts collected from one such workshop. Lens1 defines Richmond's systems thinking skills (dynamic, system as cause, closed loop, forest, operational) at the artifact level, with explicit indicators for what each skill looks like in an Iceberg model and in a CLD (Richmond, 1993). Lens2 defines opportunity identification signals (event to structure conversion, boundary expansion, feedback driven leverage, surfaced feasibility or adoption constraints), tying artifact features to constructs from the opportunity development literature (Ardichvili et al., 2003). The mapping between systems decomposition and opportunity work in Lens2 is our own

synthesis rather than a claim drawn from any single prior study; it builds on our integrative framework linking Richmond's systems thinking skills with entrepreneurial opportunity identification (Azad et al., 2024).

The two lenses are intended to be applied together: Lens1 asks whether the artifact contains evidence of systems thinking moves, Lens2 asks whether those moves are connected to identifiable intervention pathways. The protocol's contribution is methodological: it offers a behavioral, replicable instrument that complements rather than replaces survey based measures, and that can be applied without expert domain knowledge once the indicators are defined. The remainder of the paper proceeds as follows. Section 2 situates the protocol within the fragmented landscape of systems thinking assessment. Section 3 presents the protocol itself, including operational indicators, a single-page scoring rubric, and a worked annotation procedure. Section 4 describes the demonstration setting, a brief workshop with seven engineering venture teams whose artifacts are used to illustrate protocol application. Section 5 reports what the dual lens surfaced when applied to those artifacts. Section 6 discusses validity, reliability, comparisons to existing assessment approaches, and recommendations for educators. Section 7 presents the conclusion of the study.

2 Related Work: Assessment of Systems Thinking in Brief Interventions

Systems thinking assessment has long been characterized by methodological heterogeneity. A recent systematic review of assessments used in engineering identified a wide range of instruments that differ in construct definition, evidence type, scoring approach, and validation status, with no single instrument dominating the field (Dugan et al., 2022). The review found that survey based and scenario based instruments substantially outnumber artifact based ones, and recommended that the field develop assessments that examine behavioral products such as participant generated maps in addition to perceived understanding. A complementary instrument designed for complex reasoning through ill structured problems made a similar point: assessments that rely on multiple choice or rating scale formats cannot capture how learners reason about structure and feedback (Grohs et al., 2018). The picture is one of an inventory with several gaps, and the gap that matters most for brief workshops is the absence of instruments that examine what learners actually produce in the room.

Self-report is the most common evaluation strategy for short workshops, and for understandable reasons: it is inexpensive to deploy and easy to administer at scale. Its weaknesses, however, are well documented. Immediately post intervention, self-report ratings are vulnerable to response shift bias, in which respondents recalibrate the rating scale itself as a result of the experience, as well as to social desirability and demand characteristics, in which respondents infer the rating they are expected to provide (Howard & Dailey, 1979). These threats are heightened when the intervention is short, branded as a learning opportunity, and delivered by an enthusiastic facilitator, all of which describe the typical brief systems thinking workshop. None of this means self-report should be abandoned. It means it should be paired with at least one form of evidence that does not depend on the participant's own evaluative judgment.

The artifacts that brief workshops produce are well suited to play that role. Iceberg models are designed to externalize vertical decomposition from events to patterns to underlying structure and mental models (Booth Sweeney & Meadows, 2010), which means their visual structure encodes whether the participant attempted that decomposition at all. Causal loop diagrams are designed to externalize

directional causal links and closed feedback chains (Sterman, 2000), which means their visual structure encodes whether the participant attempted feedback specification. In both cases, the artifact is the participant's own record of their reasoning, not a recollection of it produced after the fact. Analyzing the artifact is therefore behavioral rather than evaluative: it scores what the participant did, not what the participant later says they understood (Grohs et al., 2018). This directly addresses the response shift and demand characteristic concerns.

Most existing artifact based assessment work in systems thinking education has been conducted in classroom or case study settings where the substantive content of the artifact is largely controlled by the instructor (Azad & Moore, 2023). Brief workshops in entrepreneurial settings break this assumption. Each team works on a different venture, in a different sector, with different stakeholders, intermediaries, and constraints. An assessment protocol intended for this setting must be interpretable across heterogeneous content. It cannot rely on the rater being an expert in any one venture's technical domain. Its indicators must instead operate on structural features of the artifact (the presence of a pattern layer, the polarity of links in a loop, the breadth of named actors) that can be recognized without venture domain knowledge.

This paper specifies and demonstrates a dual lens protocol designed for exactly this combination of constraints. Lens1 defines Richmond's systems thinking skills at the artifact level using indicators that do not require venture domain expertise. Lens2 defines opportunity identification signals using constructs from the opportunity development literature (Ardichvili et al., 2003; Clausen, 2020), combined with a leverage construct from system dynamics (Meadows, 1999); the combination is our own. The two lenses are designed to be applied together by a single rater or by independent raters whose agreement can be measured. The protocol is intended to complement, not replace, survey or focus group measures collected as part of the same evaluation. The scope of the paper is methodological: we present the protocol, illustrate its application on one workshop dataset, and discuss the reliability and validity considerations that bear on its broader adoption.

3 The Dual-Lens Assessment Protocol

This section specifies the protocol. Subsection 3.1 states the protocol's design constraints and what it is intended to assess, and fixes the vocabulary used throughout. Subsections 3.2 and 3.3 present the two lenses with operational indicators for both Iceberg models and CLDs. Subsection 3.4 consolidates both lenses into a single-page scoring rubric (Table 1) that can be applied on its own. Subsection 3.5 describes the annotation procedure and how the two lenses are applied together to a single artifact.

3.1 Design constraints and vocabulary

The protocol is intended for use on artifacts produced in brief workshops (one half day or less) with early stage engineering venture teams whose artifact content is heterogeneous in domain. Four design constraints follow from this scope. First, indicators must be recognizable on the artifact without expert knowledge of the venture's technical domain, so that a single rater or a small rater pool can score artifacts spanning a wide range of applications. Second, indicators must be observable on the artifact itself, not inferred from auxiliary materials such as transcripts or surveys, to keep the protocol independent of those data sources when they are used in parallel. Third, the protocol must accommodate the rough, incomplete artifacts typical of brief workshops: handwritten on butcher paper, partially labeled, sometimes ambiguous

in their causal direction. Fourth, the protocol must be tractable to apply: scoring a single artifact should take a competent rater on the order of ten to twenty minutes, not hours.

Because the two lenses are applied at several levels, we fix three terms and use them consistently. A *skill* (Lens1) or *signal* (Lens2) is the underlying construct being assessed, such as dynamic thinking or boundary expansion. An *indicator* is the specific, observable feature of an artifact that provides evidence for a skill or signal, such as a labeled pattern layer or a closed chain with labeled polarity. A *descriptor* is simply the written rule that defines an indicator and states what counts as absent, partial, or present. In short: a descriptor defines an indicator, and an indicator is evidence of a skill or signal. Each indicator is scored on a three-level ordinal scale — *absent*, *partial*, or *present* — rather than as a binary judgment. We adopt a partial level deliberately, because in brief workshops the modal outcome on several indicators is an incomplete reasoning move (a structure named but not closed, a pattern implied but not labeled), and a binary scheme would discard exactly the resolution the protocol is meant to provide.

3.2 Lens1: systems thinking skills

Lens1 assesses five of Richmond's systems thinking skills at the artifact level (Richmond, 1993, 1998). We developed the indicators below by taking each skill as Richmond defines it and asking what its presence would look like as a visible feature on a workshop artifact; the indicators are therefore our operationalization of Richmond's constructs, not Richmond's own categories, and they are intended to be recognizable on either an Iceberg model or a CLD (Azad et al., 2024).

For each skill we specify the indicator, what it looks like on each artifact type, and what distinguishes a present score from a partial one. The complete set of absent/partial/present rules is consolidated in the scoring rubric (Table 1).

Dynamic thinking. Indicator: the artifact frames the challenge as behavior over time rather than as a one off event. On an Iceberg model this appears as an explicit pattern or trend layer; on a CLD it appears as variables that represent quantities changing over time (e.g., rate of adoption, level of trust) rather than static labels. *Present*: time or trend framing is explicit. *Partial*: change over time is implied but not named as a recurring pattern (e.g., a single escalation without a pattern layer). *Absent*: the framing is static or purely event level.

System as cause thinking. Indicator: observed outcomes are attributed to internal system design, constraints, or shared assumptions rather than to external shocks. On an Iceberg model this appears in the structure or mental model layers; on a CLD it appears as closed chains in which the cause of a problem is located inside the variable set. *Present*: the explanation is predominantly endogenous. *Partial*: the artifact mixes endogenous and exogenous attribution. *Absent*: outcomes are attributed to external shocks.

Closed loop thinking. Indicator (CLD only): at least one closed feedback chain with labeled polarity, identifying a reinforcing or balancing loop. This skill is the one indicator in Lens1 that is not assessable on an Iceberg model, because an Iceberg has no loop structure to inspect. *Present*: a closed loop with explicit polarity. *Partial*: a closed chain without labeled polarity, or a plausibility chain that begins and ends at different variables without closure. *Absent*: only linear cause-effect chains.

Forest thinking. Indicator: boundary breadth, measured as the number of actor categories named beyond the focal user (for example, intermediaries, gatekeepers, regulators, platforms, suppliers, financiers). *Present*: at least three actor categories beyond the focal user. *Partial*: one or two such categories. *Absent*: only the focal user, the team, and the product.

Operational thinking. Indicator: variables and structural drivers are specified as mechanisms rather than as abstract labels (e.g., "verification turnaround time" rather than "delays"; "verified buyer profile coverage" rather than "trust"). *Present*: most labels are mechanism level. *Partial*: the artifact mixes mechanism level labels and placeholders. *Absent*: labels are placeholders only. Operational specificity is the indicator most directly tied to whether the artifact would support hypothesis testing in subsequent work.

3.3 Lens2: opportunity identification signals

Lens2 assesses whether the systems thinking moves in Lens1 are connected to identifiable intervention pathways. Each signal is scored on the same three-level scale (absent, partial, present).

Event to structure conversion. Indicator: the artifact contains an explicit chain that traces a surface symptom into one or more structural drivers and positions at least one candidate intervention at the structural layer. The intervention is the opportunity signal; structure alone is descriptive. *Present*: chain reaches structure and names a candidate intervention there. *Partial*: chain reaches structure but names no candidate intervention. *Absent*: no descent below the event or pattern layer. **Boundary expansion.** Indicator: the artifact includes at least one intermediary, gatekeeper, platform, regulator, or enabling infrastructure actor whose decisions or constraints would shape adoption. This signal overlaps with forest thinking in Lens1 but is scored separately because the question for Lens2 is whether the expanded boundary alters the team's stated opportunity set. *Present*: an added actor whose constraints visibly alter the stated opportunity. *Partial*: additional actors are named but no opportunity implication is articulated. *Absent*: no actors beyond the focal user.

Feedback driven leverage. Indicator: the artifact identifies at least one variable or relationship within a feedback structure as a candidate point of intervention, with an articulation of why a change at that point is conjectured to propagate through the loop (Meadows, 1999). *Present*: a leverage point named with propagation logic. *Partial*: a leverage point named without articulating its propagation. *Absent*: no leverage point identified.

Surfaced feasibility or adoption constraint. Indicator: the artifact names at least one constraint (cost, regulatory approval, trust, switching cost, distribution access) that would shape whether the opportunity is realizable. This signal captures the move from "would this be valuable" to "is this achievable." *Present*: at least one constraint named explicitly. *Partial*: a constraint alluded to but not specified. *Absent*: no constraint surfaced.

3.4 The scoring rubric

Table 1 consolidates both lenses into a single reference rubric. It is designed to be used on its own: a rater can score any system map by reading down the nine rows and selecting, for each, the column whose description best matches the artifact. Each indicator is scored absent (0), partial (1), or present (2). We deliberately do not collapse the nine indicator scores into a single total. The protocol's value lies in the *profile* of scores — which skills are present, which stop at partial, and whether Lens1 evidence is matched by Lens2 evidence — and a sum would discard exactly that resolution. A compact way to report a single artifact is therefore the count of present and partial indicators within each lens (for example, "Lens1: 3 present, 1 partial, 1 absent; Lens2: 1 present, 2 partial, 1 absent"), which preserves the distribution while remaining easy to tabulate across a cohort.

Table 1: The dual-lens scoring rubric. Score each indicator **absent (0)**, **partial (1)**, or **present (2)** by matching the artifact to the best-fitting column. Closed loop thinking is assessable only on tools that represent feedback explicitly (CLD, stock-and-flow).

Indicator	Absent (0)	Partial (1)	Present (2)
Lens1 — Systems thinking skills (Richmond)			
Dynamic thinking	Static or event-level framing only	Change over time implied but not named as a pattern	Explicit pattern/trend layer (Iceberg) or time-varying variables (CLD)
System as cause thinking	Outcomes blamed on external shocks	Mix of endogenous and exogenous attribution	Outcomes attributed to internal structure, constraints, or assumptions
Closed loop thinking *(CLD only)*	Only linear cause–effect chains	Closed chain without polarity, or an unclosed plausibility chain	At least one closed feedback loop with labeled polarity
Forest thinking	Only focal user, team, and product	One or two actor categories beyond the focal user	Three or more actor categories beyond the focal user
Operational thinking	Placeholder labels only (e.g., "delays")	Mix of mechanism-level labels and placeholders	Mostly mechanism-level variables/drivers (e.g., "turnaround time")
Lens2 — Opportunity identification signals			
Event to structure conversion	No descent below the event/pattern layer	Chain reaches structure but names no intervention	Symptom → structural driver → candidate intervention at structure
Boundary expansion	No actors beyond the focal user	Extra actors named but no opportunity implication	Added actor whose constraints alter the stated opportunity set
Feedback driven leverage	No leverage point identified	Leverage point named without propagation logic	Leverage point named *with* an articulated propagation path
Surfaced constraint	No feasibility/adoption constraint surfaced	A constraint alluded to but not specified	At least one feasibility/adoption constraint named explicitly

3.5 Annotation procedure

Annotation proceeds in four steps per artifact, using the rubric in Table 1. (1) The rater catalogues the artifact's surface features: layers present (for Icebergs), variables, links and polarities (for CLDs), and named actors. (2) The rater applies the Lens1 indicators in sequence, recording absent, partial, or present for each, with a one to two sentence justification anchored to a specific feature of the artifact. (3) The rater applies the Lens2 indicators, again with anchored justifications. (4) For each artifact, the rater notes any features that would be informative to revisit if paired survey or transcript evidence becomes available, but does not consult those sources during lens application. When multiple raters are used, scores are recorded independently before a reconciliation pass, and disagreements are logged at the indicator level rather than smoothed.

4 Demonstration Setting and Methods

This section describes the demonstration setting, a brief workshop with early-stage engineering venture teams whose artifacts are used to illustrate protocol application in Section 5. The workshop's substantive learning outcomes are reported in a parallel manuscript (Azad, 2026); here we summarize only what is needed to interpret the artifacts. The demonstration workshop was a four hour in person session for early stage engineering venture teams, structured around the Iceberg Model and causal loop diagrams. Twenty nine teams registered interest; eight were selected against pre-specified inclusion criteria (active ideation to early validation stage, working team of at least two, live venture challenge, full session availability); seven attended. Across the seven teams, 22 individual participants brought venture contexts that spanned digital platforms, hardware and deep tech, health and care technology, workforce and marketplace systems, sustainability and climate technology, and consumer applications (Table 2). The session combined short tool introductions, team mapping activities scaffolded by a single facilitator, and a closing reflection and sharing segment.

Table 2: Demonstration cohort: venture domain and stage of the seven participating teams.

Team	**Venture domain**	**Stage at workshop**
Team A	Digital platform and services	Ideation to early validation
Team B	Hardware or deep tech	Ideation to prototyping
Team C	Health or care technology	Prototyping
Team D	Workforce or marketplace system	Early validation
Team E	Sustainability or climate related technology	Ideation
Team F	Consumer product or application	Prototyping
Team G	Mixed (not included in the analysis)	Attended only

Photographs and exports of team-produced Iceberg models and causal loop diagrams were collected at the close of the session. Twelve artifacts from six teams (six Icebergs and six CLDs) form the demonstration set used in Section 5. One team did not produce a CLD within the session and is not included in the analysis. Artifacts were stored as image files with team identifiers anonymized at the file level. No transcript or survey data are used in the protocol demonstration; those data sources are part of a parallel evaluation reported separately, and the present paper deliberately confines its evidence to what the artifacts themselves contain. Anonymized exemplars are shown in Figure 1 (Section 5).

The first author, who also designed and facilitated the workshop, applied the protocol described in Section 3 to all twelve artifacts. For each artifact, the four step procedure was followed (catalogue, apply Lens1, apply Lens2, note paired evidence questions), and indicator level scores were recorded with one to two sentence justifications anchored to specific artifact features. To reduce the risk of focus group or survey content biasing the artifact scores, annotation was conducted independently of those two data streams: artifacts were scored approximately one week after the workshop, before any focus group transcripts were coded and before any survey results were analyzed, and the rater did not consult those sources during lens application. We acknowledge that this procedure is not equivalent to independent rater scoring by someone outside the research team, and we treat the demonstration as illustrative rather than as a reliability study. Independent rater agreement on this protocol has not yet been measured and is identified as a priority in Section 6.

Ethics approval was obtained from the University of Toronto Health Sciences Research Ethics Board (Protocol #47007). Informed consent was obtained from all participants, including consent for the inclusion of anonymized artifact images in publications. AI assisted tools were used only for transcription and minor language editing during preparation of interview notes for the parallel evaluation; no AI tools were involved in the protocol design, the artifact annotation, or the writing of this paper beyond minor language editing. The authors reviewed and edited all materials and take responsibility for the final content.

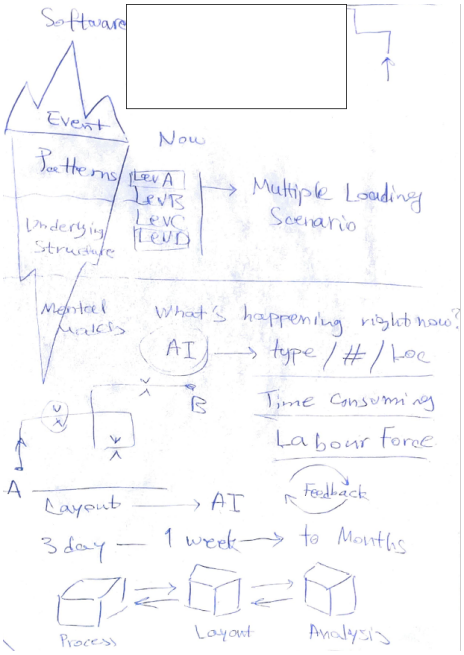
5 Applying the Protocol: Findings from the Demonstration Set

This section reports what the dual lens surfaced when applied to the twelve artifacts in the demonstration set. We organize the findings around three questions the protocol is designed to answer: which systems thinking skills are visible at the artifact level (Lens1), whether those skills are connected to identifiable intervention pathways (Lens2), and what patterns the lens scores expose across teams. We begin with a worked example on two anonymized artifacts, which shows how the indicators are applied to a single team's work, and then present the indicator distributions across the full set in Tables 3 and 4.

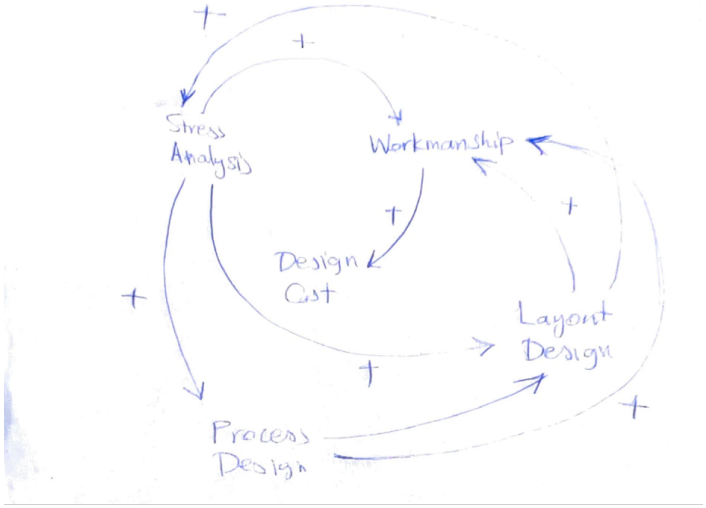
5.1 Worked example

Figure 1a shows an anonymized Iceberg from a team whose redacted upper right area contained three named third party software packages relevant to the team's venture. Applying Lens1, the artifact scores *present* on dynamic thinking (the pattern layer names multiple loading scenarios as a recurring trend), *present* on system as cause thinking (the structure layer attributes turnaround variability to internal labor force and process design constraints rather than to external shocks), *partial* on forest thinking (named actors beyond the focal user are limited to a process-to-layout-to-analysis workflow, i.e., one to two actor categories), and *partial* on operational thinking (some labels are mechanism level, others are placeholders such as "Time Consuming"). Applying Lens2, the artifact scores *present* on event to structure conversion (the days-to-months pattern is traced into structural drivers that are explicitly framed as candidate intervention targets), *partial* on boundary expansion, and *present* on surfaced constraint (labor force is named as a binding constraint).

Figure 1b shows an anonymized CLD from the same team. Applying Lens1, the artifact scores *present* on closed loop thinking (closed chains with labeled polarity among stress analysis, workmanship, design cost, layout design, and process design), *present* on system as cause thinking (the loops position the team's pain points as endogenously generated), and *partial* on operational thinking (variables are named at the construct level rather than as quantities). Applying Lens2, the artifact scores *partial* on feedback driven leverage (a candidate intervention point can be inferred from the loop structure, but the propagation logic is not explicitly articulated). Together, the two artifacts illustrate how the lenses combine: Lens1 confirms that the team executed the systems thinking moves the workshop was designed to elicit, and Lens2 clarifies which of those moves connected to identifiable opportunity work and which stopped at description.



(a) Participant-produced Iceberg model. The blank block at upper right covers three named third party software packages redacted at the request of the participating team.



(b) Participant-produced causal loop diagram.

Figure 1: Anonymized exemplars used as worked examples for protocol application.

5.2 Lens1: distribution across the set

Table 3 summarizes the Lens1 indicator distribution across the demonstration set. Because the same skills can in principle appear on either artifact type, the table reports Iceberg and CLD scores side by side rather than assigning each skill to a single artifact; closed loop thinking is the one skill that is not assessable on an Iceberg and is marked accordingly. Across the six Iceberg artifacts, dynamic thinking and system as cause thinking were the indicators most frequently present, observed at least partially in five of six artifacts. Forest thinking was present in four of six (counting artifacts that named at least three actor categories beyond the focal user). Operational thinking was the most variable indicator, ranging from concrete, mechanism level labels (present) to placeholder language that nominally occupied the structure layer without specifying a mechanism (absent). On the CLDs, closed loop thinking was present in four of six artifacts with labeled polarity, partial in one (closed chain without explicit polarity), and absent in one (linear plausibility chain). Scored on the CLDs, the remaining skills distributed as follows: dynamic thinking was present in the two artifacts whose variables were expressed as time-varying quantities (for example, demand, availability, and adoption) and absent in three whose nodes named static actors; system as cause thinking was present in three artifacts whose loops located causation inside the variable set; and forest thinking was present in two artifacts that named three or more actor categories. Reading the Iceberg and CLD columns side by side makes visible that a team could score differently on the same skill across its two artifacts, which is information a single-artifact analysis would miss.

Table 3: Lens1 (systems thinking skills) indicator distribution across the demonstration set ($n = 6$ Iceberg artifacts; $n = 6$ CLD artifacts). Cells give the number of artifacts scored present (P), partial (X), or absent (A) on each artifact type. "–" marks a skill not assessable on that artifact type.

	Iceberg ($n = 6$)			**CLD** ($n = 6$)			
Skill	**P**	**X**	**A**	**P**	**X**	**A**	**Indicator**
Dynamic thinking	4	1	1	2	1	3	Pattern/trend layer; time-varying variables
System as cause thinking	3	2	1	3	2	1	Endogenous explanation; closed causal chain
Closed loop thinking	–	–	–	4	1	1	Closed chain with labeled polarity
Forest thinking	4	1	1	2	3	1	≥ 3 actor categories beyond focal user
Operational thinking	2	2	2	3	2	1	Mechanism-level labels vs. placeholders

5.3 Lens2: distribution across the set

Lens2 scores were generally lower than Lens1 scores in the demonstration set, which is consistent with the design hypothesis underlying the lens: it asks whether systems thinking moves are connected to identifiable intervention pathways, a more demanding criterion than producing the structural decomposition alone. Table 4 summarizes the Lens2 distribution. Event to structure conversion was present in three of six artifacts, partial in two, and absent in one. Boundary expansion at the opportunity level was present in three of six and partial in the remaining three; the same artifact could score on forest thinking in Lens1 without scoring as present on boundary expansion in Lens2 when the additional actors were named but no opportunity implication was articulated. Feedback driven leverage was present in two of six CLD artifacts, partial in one, and absent in three. Surfaced feasibility or adoption constraint was present in four of six artifacts and partial in two, making it the most frequently observed Lens2 signal.

Table 4: Lens2 (opportunity identification signals) distribution across the demonstration set ($n = 6$ teams). Counts give the number of artifacts scored present (P), partial (X), or absent (A) on each signal.

Signal	**P**	**X**	**A**	**Indicator**
Event to structure conversion	3	2	1	Symptom $\rightarrow$ structural driver $\rightarrow$ candidate intervention
Boundary expansion	3	3	0	Added actor that alters the stated opportunity set
Feedback driven leverage	2	1	3	Leverage point named with propagation logic
Surfaced constraint	4	2	0	Feasibility/adoption constraint named explicitly

5.4 Cross-team patterns

The protocol's value at Lens2 is not the absolute count but the differentiation it produces. Three artifacts scored present on at least three of four Lens2 signals; these are the artifacts on which the workshop produced not just a systems decomposition but a connected pathway from structure to candidate intervention. Two artifacts scored present on at most one Lens2 signal; these are the artifacts on which a systems decomposition was produced but did not translate into identifiable opportunity work. An aggregate self-report measure cannot distinguish between these two groups. Two observations are worth flagging because they are the kind of pattern the protocol is built to surface. First, partial scores are common and informative. The partial level is not a fudge: it captures artifacts that demonstrate the structure of a systems thinking move without completing it. In a brief workshop, partial evidence is

the modal outcome on several indicators, and a binary present-or-absent scheme (which several survey based instruments effectively reduce to) loses this resolution. Second, Lens1 and Lens2 can dissociate. An artifact can score well on Lens1 and poorly on Lens2 when teams produce competent structural decompositions but do not connect those decompositions to candidate interventions; this is the pattern most relevant to entrepreneurship educators, because it identifies teams who have learned the tool without yet using it for opportunity work. The asymmetry across artifacts is itself informative: brief workshops produce a non uniform distribution of indicators, which an aggregate self-report measure would compress into a single perceived understanding score and which the protocol instead surfaces as a distribution available for inspection.

6 Discussion

6.1 What the protocol contributes to systems thinking assessment

The dual lens protocol contributes a behavioral, replicable instrument for assessing systems thinking in brief workshops where self-report measures are most vulnerable to bias. Two design features distinguish it from the survey based and scenario based instruments catalogued in recent reviews (Dugan et al., 2022; Grohs et al., 2018). First, it operates directly on the artifacts the workshop already produces, requiring no additional participant burden beyond the workshop itself. Second, it ties artifact features to constructs from two established literatures (Richmond's systems thinking skills and the opportunity identification tradition), so that scores are interpretable in terms of recognizable theoretical anchors rather than as bespoke rubric categories that future researchers would need to learn to compare across studies. The protocol does not solve the deeper problem of assessment fragmentation in systems thinking education; it adds one more instrument to the inventory. Its claim is to be a useful addition because it covers an under served combination: short interventions, behavioral evidence, and heterogeneous artifact content.

6.2 The lenses are not unique to the Iceberg and CLD

Neither lens is intrinsically tied to the two tools used in this demonstration. The constructs the lenses assess — Richmond's systems thinking skills and the opportunity identification signals — are learning outcomes that could surface in many representations: a stock-and-flow diagram, a behavior-over-time graph, a stakeholder or rich-picture map, a written venture memo, or a verbal pitch could each carry evidence of dynamic thinking, endogenous causal reasoning, boundary breadth, or a named leverage point. What the lenses assess is the reasoning move, not the diagram. We focused on the Iceberg model and the CLD for a specific reason: our prior framework explicitly maps Richmond's systems thinking skills to the representational affordances of these two tools, establishing which skills each tool is best positioned to elicit and make visible (Azad et al., 2024). That mapping is what lets us write indicators that are confident, tool-specific, and quick to apply. It is not a claim that these tools have a monopoly on the underlying outcomes. The same set of measures can therefore be extended beyond the tools demonstrated here, provided the indicators are re-specified for the new representation, and we regard such extension as a feature of the protocol rather than a limitation of it.

6.3 Transferability to other diagram types

The distinction we draw between skills and indicators makes the path to extension concrete. Richmond's framework is tool independent: the five systems thinking skills are the constructs being assessed regardless of which mapping tool a workshop uses (Richmond, 1993, 1998). What is tool dependent is the set of indicators, because each indicator is defined in terms of a visible feature of a particular artifact type. Lens1 should therefore transfer to other system dynamics tools with re-specified indicators rather than new constructs. A stock and flow diagram, for example, would express dynamic thinking through accumulations and rates rather than a pattern layer, and operational thinking through the explicitness of rate equations; a behavior over time graph would express dynamic thinking directly and closed loop thinking not at all. The tool-to-skill mapping in our prior framework (Azad et al., 2024) provides a starting point for porting the indicators: a tool elicits a skill well when its native representation makes the corresponding indicator visible. Closed loop thinking is the clearest boundary case: it is assessable only on tools that represent feedback explicitly (CLDs, stock and flow), which is why it is marked not assessable on the Iceberg in Table 3. Lens2, being defined on actors, boundaries, and intervention points rather than on a specific diagrammatic convention, is less tool bound than Lens1, though the feedback driven leverage signal inherits the same dependence on a tool that represents loops. We expect the protocol to be most robust within the family of tools for which it was designed and to require indicator recalibration, not reconstruction, for substantially different artifact types.

6.4 Validity and reliability considerations

Construct validity for Lens1 rests on whether the indicators correctly capture Richmond's systems thinking skills at the artifact level. The indicators provided are deliberately conservative by targeting features that are visually inspectable on a typical workshop artifact (a labeled pattern layer, polarity on links, count of actor categories) rather than features that would require interpretation of the team's reasoning process. This trades sensitivity for replicability. A more sensitive instrument might score artifacts on the quality of their feedback hypotheses; the protocol as specified scores only the presence of closed chains with polarity.

Researchers who need a more graduated measure could extend the Lens1 indicators without breaking the protocol's overall structure. Construct validity for Lens2 is similar but pivots on whether the signals correctly capture connection from systems decomposition to identifiable intervention pathways. The four signals selected (event to structure conversion, boundary expansion at the opportunity level, feedback driven leverage, surfaced constraint) draw from opportunity development theory (Ardichvili et al., 2003) combined with a leverage construct from system dynamics (Meadows, 1999), but they are not exhaustive. Researchers operating in venture domains where regulatory or institutional dynamics dominate (health, energy, public sector) may want to add a signal for institutional pathway specification; researchers in consumer domains may want to add one for adoption mechanism specification. The protocol's structure accommodates this kind of extension because the signals are organized as a list rather than as an integrated score.

The protocol's reliability has not yet been quantified. The demonstration set was scored by the first author, who also designed and facilitated the workshop; that familiarity with the workshop context could plausibly inflate agreement with the indicators, even though scoring was conducted independently of the focus group and survey strands. The minimum reliability evidence required for adoption is independent

rater agreement at the indicator level by raters outside the research team, ideally reported as a percent agreement (Krippendorff, 2004).

Two additional reliability questions matter for educators who want to use the protocol in their own programs: whether indicators are robust to variation in artifact medium (handwritten on butcher paper vs. produced in mapping software), and whether they transfer across domains (entrepreneurial venture artifacts vs. classroom case artifacts vs. organizational problem artifacts). We expect the protocol to be most robust in the entrepreneurial range for which it was designed and to require recalibration of the forest thinking and operational thinking indicators for substantially different artifact contexts.

6.5 Relationship to other assessment approaches

The protocol is intended to complement, not replace, survey based instruments. Surveys remain the most efficient way to measure perceived understanding and perceived relevance, and they capture participant judgments that artifact analysis cannot. The protocol's role is to provide a parallel evidence stream that does not depend on participant evaluative judgment, so that paired analysis can detect cases in which self rated understanding rises without artifact evidence of systems thinking moves (a likely sign of response shift or demand effect) or cases in which artifact evidence is strong but self rated understanding does not move (a likely sign of imposter or recalibration effects). Similarly, the protocol complements rather than replaces scenario based instruments such as those built around ill structured problems (Grohs et al., 2018): scenarios standardize content and enable cross study comparison; artifacts retain the situated reasoning that motivates entrepreneurship education to use authentic venture contexts in the first place. The protocol is positioned to be useful where authentic content is non-negotiable and self-report is insufficient.

6.6 Recommendations for educators

For educators planning to use the protocol, three practical recommendations follow from the demonstration. First, plan to collect artifacts in a form that supports later annotation: legible photographs taken at the close of the session, with the team's name written by the team at a designated corner that can be cropped for anonymization. Second, schedule annotation within one to two weeks of the session while facilitator memory of the room is still fresh; the protocol does not require facilitator notes, but recent context helps the rater avoid over interpreting ambiguous features. Third, pair the protocol with at least one other evidence stream, typically a short self-report measure on perceived understanding and relevance, so that the convergent and divergent cases described above become detectable. The rubric in Table 1 is designed to travel with the artifacts: it can be printed alongside the collected maps and scored directly, requiring no software and no venture domain expertise.

6.7 Limitations

The protocol has not yet been validated by independent rater agreement; the demonstration set is small and from a single workshop in a single domain mix; the rater was the first author, who also designed and facilitated the workshop and was therefore not blind to its design, although artifact scoring was conducted independently of the focus group and survey data streams; and the protocol does not include an active control comparison against artifacts produced under a non systems thinking intervention. Each of these is a priority for future work. We also note that the protocol assumes the workshop produces

both Iceberg and CLD artifacts; protocols built around other combinations of systems thinking tools (e.g., stock and flow diagrams, behavior over time graphs) would require the indicator re-specification discussed in Sections 6.2 and 6.3.

6.8 Future work

Three next steps would strengthen the protocol's evidence base. (1) An interrater reliability study with three or more independent raters scoring the same artifact set, reporting indicator level agreement statistics. (2) An active control comparison in which artifacts produced under a matched dose alternative workshop (business model canvas, design thinking) are scored using the same protocol; this would test whether the lens indicators discriminate between systems thinking interventions and non systems thinking interventions, a key validity claim. (3) A longitudinal extension that scores artifacts produced by the same teams at multiple timepoints, allowing the protocol to capture trajectory rather than only cross sectional level.

7 Conclusion

This paper has presented a dual lens protocol for assessing systems thinking artifacts produced in brief workshops with early stage engineering venture teams. The protocol's contribution is methodological: it specifies indicators operating on artifact features that are recognizable without expert knowledge of the venture's technical domain, ties those indicators to constructs from established systems thinking and opportunity identification literatures, and supports application by a single rater or by independent raters whose agreement can be measured. Applied to a demonstration set of twelve artifacts from one workshop, the protocol distinguished artifacts that completed structural decomposition from those that did not, identified which teams reached the leverage hypothesis stage, and exposed indicator level variation that aggregate self-report measures would compress into a single score.

The protocol is offered as a complement to, not a replacement for, survey and scenario based instruments. Its appropriate use is in contexts where artifact based behavioral evidence is needed alongside or in place of self-report, particularly in brief interventions where self-report is most vulnerable to bias and where artifact content is too heterogeneous for instructor controlled scenario instruments. The next steps that would most strengthen the protocol are independent rater reliability, an active control comparison against non systems thinking interventions of matched dose, and longitudinal application across repeated workshop sittings.

Author Contributions

Amin Azad: Conceptualization, Methodology, Investigation, Formal Analysis, Writing – Original Draft. **Emily Moore:** Supervision, Writing – Review & Editing.

Funding

This research was funded by the NEO Materials Catalyst Fund. Workshop materials and facilitation were supported in kind by the Institute for Studies in Transdisciplinary Engineering Education and Practice (ISTEP), University of Toronto.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Anonymized aggregate data and the workshop curriculum are available from the corresponding author on reasonable request, subject to participant consent constraints.

Ethics Approval Statement

This study was conducted in accordance with the Declaration of Helsinki and approved by the University of Toronto Health Sciences Research Ethics Board (Protocol #47007). Informed consent was obtained from all participants prior to data collection.

Acknowledgments

The authors gratefully acknowledge the engineering venture teams who participated in the workshop and shared their mapping artifacts for analysis.

References

Ardichvili, A., Cardozo, R., & Ray, S. (2003). A theory of entrepreneurial opportunity identification and development. *Journal of Business Venturing*, *18*(1), 105–123.

Azad, A., Moore, E., & Hounsell, A. (2024). Advancing engineering education: Linking systems thinking skills to the tools through a revised framework. In *2024 ASEE Annual Conference & Exposition Proceedings*. Portland, OR. https://doi.org/10.18260/1-2--46525

Azad, A., & Moore, E. (2023). Landscape review of entrepreneurship education in Canada and the presence of systems thinking. In *2023 ASEE Annual Conference & Exposition Proceedings*. Baltimore, MD. https://doi.org/10.18260/1-2--43932

Azad, A. (2026). *[Systems thinking for opportunity identification in engineering entrepreneurship education, Ph.D. Thesis, University of Toronto]*.

Booth Sweeney, L., & Meadows, D. H. (2010). *The systems thinking playbook*. Chelsea Green.

Clausen, T. H. (2020). Entrepreneurial thinking and action in opportunity development: A conceptual process model. *International Small Business Journal*, *38*(1), 21–40.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Dugan, K. E., Mosyjowski, E., Daly, S., & Lattuca, L. (2022). Systems thinking assessments in engineering: A systematic literature review. *Systems Research and Behavioral Science*, *39*(4), 840–866.

Grohs, J. R., Kirk, G. R., Soledad, M. M., & Knight, D. B. (2018). Assessing systems thinking: A tool to measure complex reasoning through ill-structured problems. *Thinking Skills and Creativity*, *28*, 110–130.

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*(2), 144–150.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Meadows, D. H. (1999). *Leverage points: Places to intervene in a system*. Sustainability Institute.

Richmond, B. (1993). Systems thinking: Critical thinking skills for the 1990s and beyond. *System Dynamics Review*, *9*(2), 113–133.

Richmond, B. (1998). The thinking in systems thinking: How can we make it easier to master. *The Systems Thinker*, *9*(2), 1–5.

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization*. Doubleday.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology*, *74*(2), 265–272.

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