DECISION SUPPORT LABS

Decision Lab™

Data Architecture, Simulation Methods, and Validation Roadmap - Version II

Methods Paper v0.2 | Version II | September 2026

A school-level simulation framework for comparing intervention strategies before implementation - with explicit assumptions, multi-semester student response, district economics, and a prospective validation loop.

Research-stage technical paper. Scenario outputs are modeled planning estimates, not causal findings or guaranteed forecasts.

Prepared for district, research, technology, and strategic review

High School Decision Support • DecisionSupportLabs.com

Abstract

Decision Lab is a high-school decision-support environment designed to compare plausible intervention strategies before a district commits staff time, budget, or implementation capacity. The system combines school baselines, official public data, longitudinal research, local assumptions, and - as the Campus Engagement System is deployed - school-authorized behavioral and intervention-response signals.

The experimental research line represents a synthetic high-school population at the student level and carries that population across multiple semesters. Rather than treating each semester as a fresh start, students can begin support, complete or stop it, continue or fade, respond differently, and accumulate effects over time. The aim is not to predict a named student's future. The aim is to help leaders compare choices under explicit assumptions and then learn from what actually happens.

This paper documents the data architecture, synthetic-population logic, intervention-response mechanics, multi-semester transition model, output construction, and validation roadmap. Version II extends the architecture in five directions: observed behavior and intervention-response data, experimentally stronger local evidence, peer and network effects, direct benchmarking against simpler forecasts and predictive models, and a confidence layer tied to held-out error. Decision Lab should be evaluated as a planning simulation until prospective school implementations establish the conditions under which its modeled effects are sufficiently calibrated for stronger predictive claims.

Eight design principles

1. School-specific	Start from the selected school and district context, not a generic national average.
2. Behavioral	Model what students actually do: participation, follow-through, stopping, continuation, response, and fade.
3. Intervention-aware	Represent what support was delivered, to whom, when, at what intensity, and what happened afterward.
4. Comparative	Use the same baseline to compare strategy mixes, staffing assumptions, constraints, and horizons.
5. Transparent	Separate observed data, research anchors, educator inputs, user assumptions, and modeled outputs.
6. Benchmarkable	Compare the simulation against simple forecasts and predictive baselines on the same held-out future outcomes.
7. Confidence-aware	Report uncertainty and, after validation, scenario-level confidence tied to observed model error.
8. Learnable	Close the loop by comparing modeled vs. actual results and recalibrating the next simulation.

Related research context. Version II is informed by three lessons: richer person-level grounding can improve simulation relative to demographic-only descriptions [7]; training on experimental-response data can improve generalization to unseen studies and conditions [8]; and decision-grade simulation requires held-out validation and confidence estimates rather than average accuracy alone [12][13]. These findings motivate the architecture but do not validate Decision Lab.

What Version II adds from recent behavioral-simulation research

External finding	Design implication for Decision Lab
Richer person-level grounding can outperform demographic-only descriptions, but reported accuracy should be interpreted relative to the benchmark used [7].	Preserve heterogeneous student histories and avoid treating a school average or demographic profile as a student model.
Fine-tuning on millions of experimental responses improved prediction on held-out studies and conditions [8].	Treat intervention-response data as a high-value calibration asset; do not rely only on observational correlations or general-purpose model priors.
Current LLMs can perform poorly in complex social-behavior benchmarks, with substantial subgroup variation [14].	Domain-specific data, subgroup error checks, and out-of-sample evaluation are mandatory before stronger claims.
Simile publicly emphasizes repeated validation and question-level confidence, not just average model accuracy [12][13].	Add a scenario-level confidence roadmap and measure whether “high confidence” actually predicts lower error.

1. The analytical question: not just “what happens next?”

Forecasting and simulation overlap, but they answer different planning questions. A forecast usually asks what is likely to happen if the current data-generating process continues. Decision Lab is intended for the next question: what could change if the school selects a different mix of supports, outreach, staffing effort, timing, and follow-through?

The system therefore treats a scenario as a set of explicit choices applied to a common baseline. The value of the comparison depends less on producing a single precise number than on making the assumptions visible, showing which assumptions drive the result, and identifying which strategy remains attractive when participation, completion, attrition, or response is weaker than expected.

This is consistent with analytical-quality guidance that distinguishes verification (did the model implement its specification correctly?) from validation (is the model fit for the intended decision and use environment?), and that requires uncertainty to be documented rather than hidden [10].

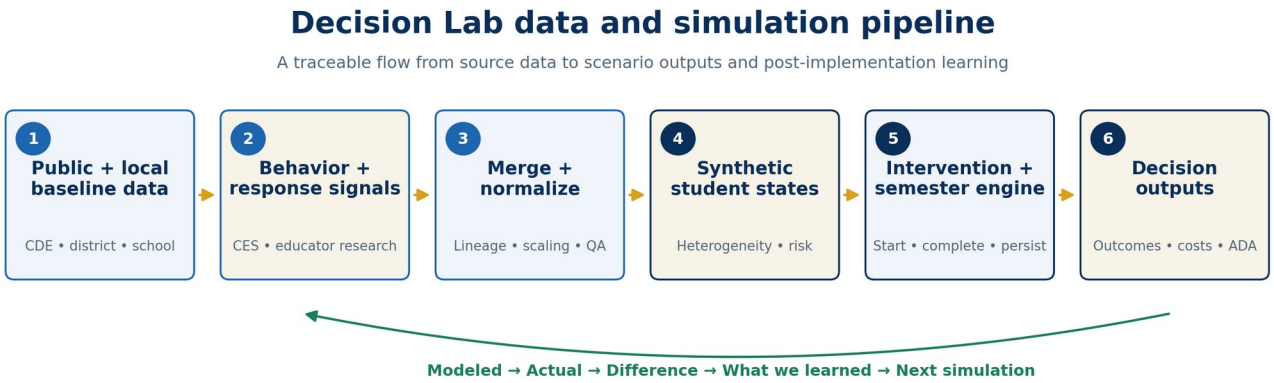

Figure 1. Decision Lab data lineage and simulation pipeline. The feedback loop is a design requirement: actual implementation results are intended to become evidence for the next model run.

2. Data foundation and lineage

Decision Lab separates five types of information because they carry different evidentiary weight:

Layer	Examples	Use in the model	Evidence status
Official public data	CDE school directory, enrollment, chronic absence, CCI	School baseline, benchmarks, selectors, historical context	Observed public data
Longitudinal research	ELS:2002, HSLS:09	Behavioral relationships, plausible distributions, prior structure	Research anchor; not local causal effect
District / school data	SIS, LMS, attendance, course and intervention records	Local baseline, implementation exposure, post-run actuals	Observed local data when authorized
Campus Engagement System	Planner acknowledgements, homework routines, participation, seconding, family visibility, intervention response	Near-real-time behavioral and response signals	Planned/ operational as systems are deployed
Experimental / quasi-experimental intervention data	Randomized encouragement, phased rollouts, timing/dosage variation	Estimate local response functions; test causal mechanisms; calibrate start, completion, persistence, and effect	Stronger evidence when design and implementation support causal interpretation

2.1 Official California baseline data

California publishes school- and district-level chronic absenteeism data and college/career data files. These sources provide official outcome baselines and comparison points for school-level scenarios [1][2]. DataQuest's 2024-25 end-of-year release also reports a statewide average of 12.8 days absent, underscoring why average days absent and chronic absence should be treated as complementary measures rather than substitutes [3].

A baseline value is not an intervention effect. Public data identify the school's starting position and comparison context; they do not establish what would happen if a particular intervention were introduced.

2.2 Longitudinal research as a prior, not a promise

ELS:2002 and HSLS:09 are nationally representative longitudinal studies that follow students across high school and into postsecondary education and work. They include student, parent, teacher, counselor, administrator, transcript, and outcome information [4][5]. Decision Lab uses relationships from these data as research anchors for the synthetic baseline and behavioral structure.

The engine does not treat a relationship observed in a national longitudinal dataset as proof that an intervention will cause the same change at a selected California school. Research relationships inform plausible model structure; local implementation data are needed to calibrate local effects.

3. Synthetic student population

The experimental microsimulation line tracks 2,000 synthetic students across the selected horizon. Each synthetic record is a model entity representing a plausible combination of attendance, academic, engagement, participation, family-visibility, and intervention-response characteristics. Synthetic students are not digital replicas of named students and are not intended to identify individuals.

The purpose of the synthetic population is to preserve heterogeneity. Two schools can have the same average attendance rate while having very different distributions of chronic absence, missing work, engagement, family visibility, and response to intervention. A school-level average alone cannot represent those differences.

The reference population is calibrated so aggregate starting values remain consistent with the selected school or reference-school baseline. School enrollment can then be used to scale aggregate outputs while the experimental engine retains a stable synthetic population size for comparability across scenarios.

3.1 Student state vector

The current research architecture organizes student state into eight variable families:

Attendance	absence rate; trajectory; risk band; tardy
Academic / LMS	missing work; on-time submission; trend; workload
Planner / peer	consistency; awareness; contribution; verification; influence
Homework	frequency; consistency; effort; responsiveness
Participation / family	intensity; breadth; activity; recognition; family visibility
Intervention	exposure; start; completion; continuation; stopping; response
Peer / network context	squad membership; peer ties; influence; exposure to peers using support
Intervention history	offer; dosage; timing; sequence; start/stop; completion; time since support; prior response

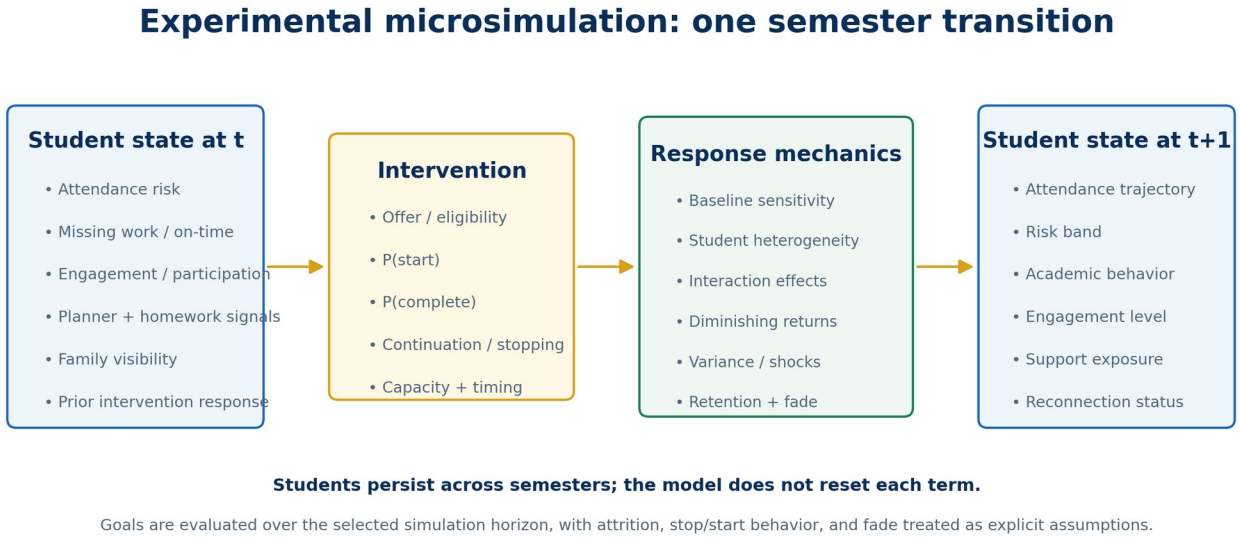

Figure 2. Conceptual semester transition in the experimental microsimulation. Exact coefficients are version-controlled model parameters and are subject to calibration.

4. Intervention-response mechanics

Decision Lab does not assume that every eligible student starts an intervention, that every starter completes it, or that an effect persists indefinitely. The model separates the intervention pathway into stages so assumptions can be tested individually.

For an intervention j applied to student i in semester t, the conceptual sequence is: eligible → offered → starts → completes or stops → responds → continues or fades. Each transition can vary by baseline risk, prior behavior, intervention fit, capacity, timing, and a stochastic response term.

$P(start_{ijt})$ = f(baseline state, eligibility, engagement, prior response, outreach, capacity)

$P(complete_{ijt} \mid start)$ = g(follow-through, workload, support intensity, friction, prior completion)

$\Delta\ state_{it}$ = h(intervention exposure, response, interactions, diminishing returns, variance)

$state_{i,t+1} = state_{it} + \Delta\ state_{it} - fade_{it} + shock_{it}$

4.1 Interaction effects

The reason to simulate rather than simply add effect sizes is that supports can interact. Peer support may increase the probability that a student uses homework support; family communication may increase follow-through; tutoring may have less incremental value when missing-work recovery has already saturated; and outreach capacity can constrain the number of students who actually receive the designed intervention.

Interaction terms should therefore be treated as assumptions until supported by research or implementation data. The model is designed so that those assumptions can be changed, sensitivity-tested, and later replaced by observed response patterns.

4.2 Reconnection as a separate pathway

Reconnection is modeled separately from attendance recovery to prevent double counting. A district-resident learner educated elsewhere can contribute new enrollment and attendance only after a modeled return, enrollment, and attendance pathway. A currently enrolled student's recovered attendance days remain in the attendance-recovery pathway. The two values may be reported together only after the model preserves that distinction.

4.3 Dosage, timing, and action sequences

Version II treats an intervention as a time-stamped exposure rather than a generic binary flag. A support can vary by eligibility, offer, dosage, timing, staff capacity, channel, sequence, and the time elapsed since the prior contact. This matters because the same support may have different effects when delivered early versus late, once versus repeatedly, alone versus after another support, or when a student has already disengaged. The student state therefore carries intervention history forward across semesters.

The conceptual target is a sequence model: school action → student starts or ignores → student completes, stops, or re-enters → short-run response → retained or fading effect → next action. This structure makes it possible to test not only “which intervention?” but also “in what order, at what intensity, and for which students?”

4.4 Behavioral grounding: what students do and why

A central Version II principle is to distinguish stated intention from observed behavior. Surveys and educator judgments can explain context and motivation, but they do not always predict what students actually do. The Campus Engagement System is therefore valuable because it can record school-authorized behavioral signals such as acknowledgement, participation, follow-through, homework routines, peer activity, intervention exposure, and continuation. These signals can be paired with educator and family context rather than replacing it.

The relevant lesson from recent behavioral-simulation research is the value of combining “what people do” with “why they say they do it.” Simile describes this as closing the say-do gap by bringing observed behavior together with interviews, surveys, and experiments [12]. Decision Lab adopts the principle, not Simile's proprietary implementation: CES behavior, educator research, and local outcome records should become complementary evidence streams.

4.5 Experimental learning stream

Where ethical, operationally feasible, and approved by the district, Decision Lab should learn from small experiments or quasi-experiments rather than only from correlations. Examples include randomized encouragement, phased rollout, or controlled variation in reminder timing, message framing, peer versus adult outreach, support intensity, or parent visibility - while never withholding legally required or educationally necessary services.

For each experiment, the system should capture assignment or rollout rule, intervention delivered, dosage, start, completion, persistence, and outcome. This creates stronger local evidence about response mechanisms and provides a defensible pathway for replacing exploratory assumptions with observed intervention-response functions.

5. Multi-semester simulation engine

In the experimental line, leaders select a simulation horizon of 2, 4, or 6 semesters. Goals are interpreted as end-of-horizon targets; the model does not reset the goal or the student population at the start of each semester.

Each semester updates the synthetic student state, intervention exposure, completion status, continuation, attrition, and any retained or fading effect. The resulting population becomes the starting point for the next semester. This allows the engine to represent momentum, delayed effects, stopping, re-entry, and the compounding consequences of different strategy sequences.

The model can therefore compare Strategy A and Strategy B under the same baseline while changing only selected assumptions. A credible comparison preserves the same random seed structure or matched stochastic runs where appropriate, so the observed difference reflects the strategy change rather than unrelated simulation noise.

5.1 Scenario parameters exposed for testing

Target horizon	2, 4, or 6 semesters
Participation / start	fraction of eligible students who begin
Completion / follow-through	fraction of starters who complete the support cycle
Continuation / attrition	likelihood of remaining active in subsequent semesters
Retention / fade	how much an effect persists after support
Response variance	heterogeneity around the average response
Capacity constraints	staff time, outreach limits, budget, or implementation intensity
Reconnection assumptions	eligible resident pool, contact/reconnect rate, persistence
Dosage / timing	frequency, intensity, delay, and sequencing of support
Peer / network influence	how exposure to participating peers changes start, persistence, or response probabilities
Comparator baseline	simple forecast and/or predictive risk model used as a benchmark
Confidence inputs	data freshness, evidence strength, local sample size, subgroup support, scenario robustness

5.2 Output families

Scenario outputs are reported at the school or district planning level. Current output families include average attendance, chronic absence, meaningful engagement, college & career preparedness, reconnected learners, days recovered, ADA-related value, cost/time burden, and strategy comparisons. Output precision should never imply more certainty than the inputs support.

5.3 Persistent state, peer effects, and multi-agent extension

Students are not independent units. Peer-to-Peer Planner squads, seconding, influencer recognition, clubs, tutoring groups, and other social structures can create spillovers: one student's participation may change another student's probability of starting or persisting. Version II therefore treats peer influence as an explicit modeling target rather than assuming that each student responds in isolation.

The current experimental engine can represent heterogeneous response and aggregate peer-influence assumptions. A fuller network model is a roadmap item, not a current validation claim. The future multi-agent extension should preserve a graph of selected peer relationships, update exposure as peers participate or stop, and test whether network-aware models improve held-out prediction without compromising privacy.

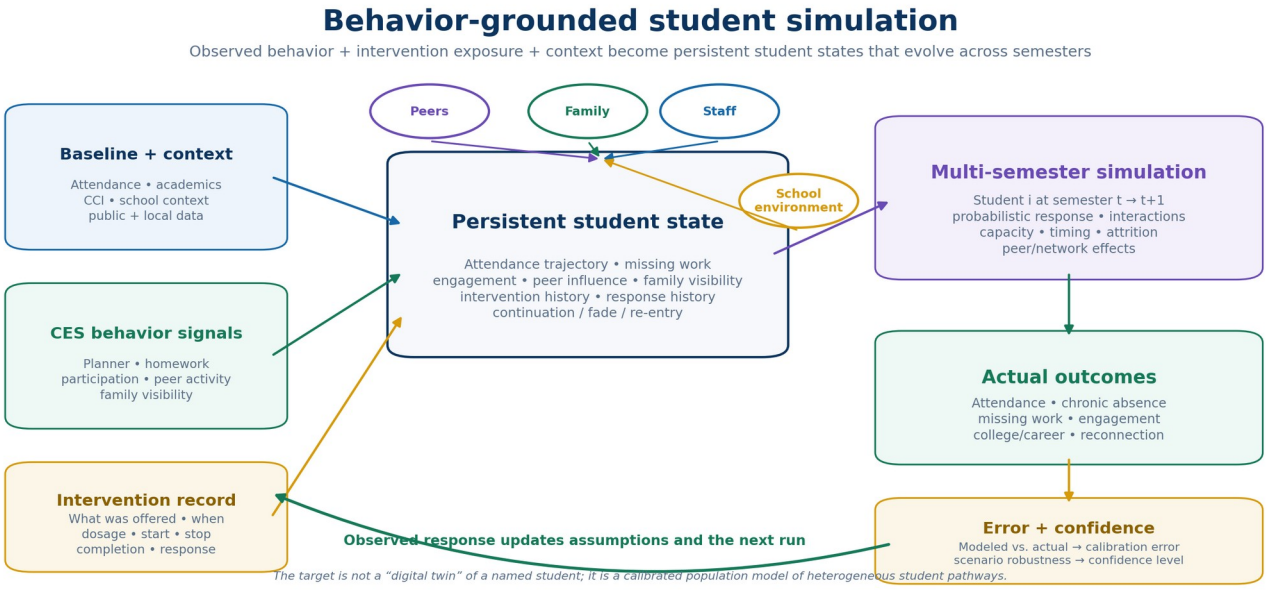

Figure 4. Version II behavior-grounded architecture. Observed CES behavior and intervention histories update persistent synthetic student states; real outcomes close the calibration loop. Peer/network effects are a development target and require validation.

6. Validation, assurance, and confidence

The central credibility question is not whether a model can produce a number; it is whether the number is fit for the decision being made. Decision Lab therefore separates software verification, benchmark calibration, assumption testing, and prospective validation.

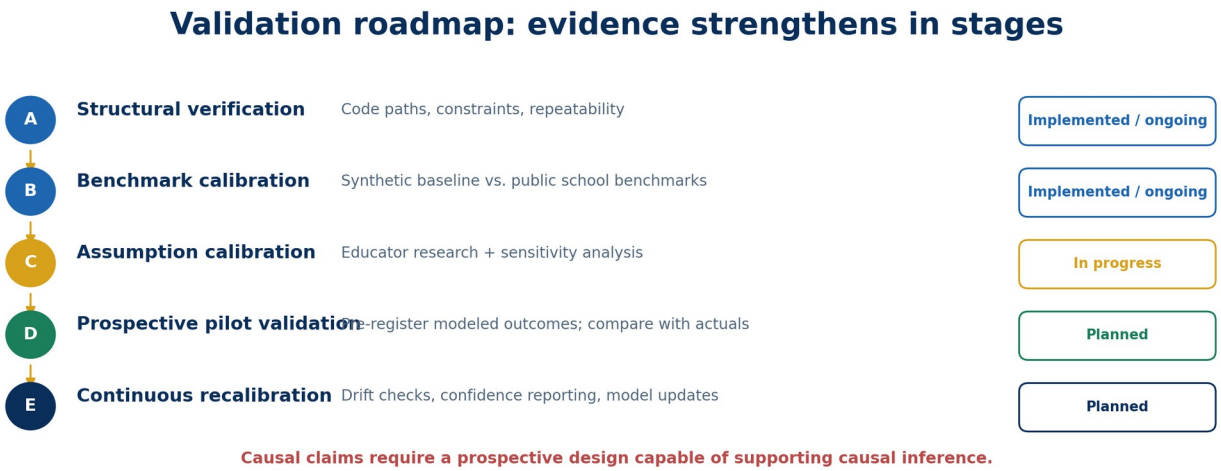

Figure 3. Staged validation roadmap. The current public and experimental systems should be described as planning simulations while prospective school validation remains outstanding.

6.1 Internal verification

Verification asks whether the software implements the intended rules. Required checks include deterministic test cases, boundary conditions, accounting identities, no-double-counting tests, scenario reproducibility, range checks, and regression tests across model versions. Version control should preserve the frozen public line separately from experimental branches.

6.2 Benchmark calibration

Calibration asks whether synthetic starting conditions reproduce the selected school or reference-school baseline and whether distributions remain plausible relative to public California data and longitudinal research. Calibration is not evidence that modeled intervention effects are causal.

6.3 Sensitivity and uncertainty

Every decision-critical output should be tested against weaker participation, lower completion, faster fade, higher attrition, capacity constraints, and alternative reconnect rates. A robust recommendation is one that remains useful across a plausible range of assumptions. Scenario ranges should be labeled as scenario ranges unless a formal statistical interval has been estimated.

6.4 Prospective validation

The decisive next step is a real-school pilot in which the baseline, intervention plan, assumptions, and modeled outcome ranges are fixed before implementation. Actual delivery and outcomes are then observed at 30, 60, and 90 days and at semester end. The comparison is: Modeled → Actual → Difference → What we learned → Next simulation. Stronger causal claims require an evaluation design capable of supporting them.

6.5 Benchmark simulation against simpler alternatives

Decision Lab should not earn credibility merely by producing more detailed scenarios. It should demonstrate incremental value over simpler methods. Every prospective validation should therefore run three approaches from the same baseline: (1) a simple trend forecast, (2) a predictive or risk model that

estimates who is most likely to experience the outcome, and (3) the intervention-aware Decision Lab simulation. All three should be frozen before the future semester is observed.

The comparison should ask two different questions. Predictive accuracy: which approach best predicts the observed aggregate and subgroup outcomes? Decision utility: which approach best ranks the strategies that later perform well under real implementation constraints? Simulation adds value only if it improves one or both of those tasks enough to matter for decisions.

6.6 Held-out evaluation

A credible validation period must be held out. The model, coefficients, scenario choices, and forecast comparator are frozen before the outcome window. After the semester, the same observed outcomes are scored against each approach. This prevents retrospective tuning from being mistaken for prediction and creates an auditable record of out-of-sample performance.

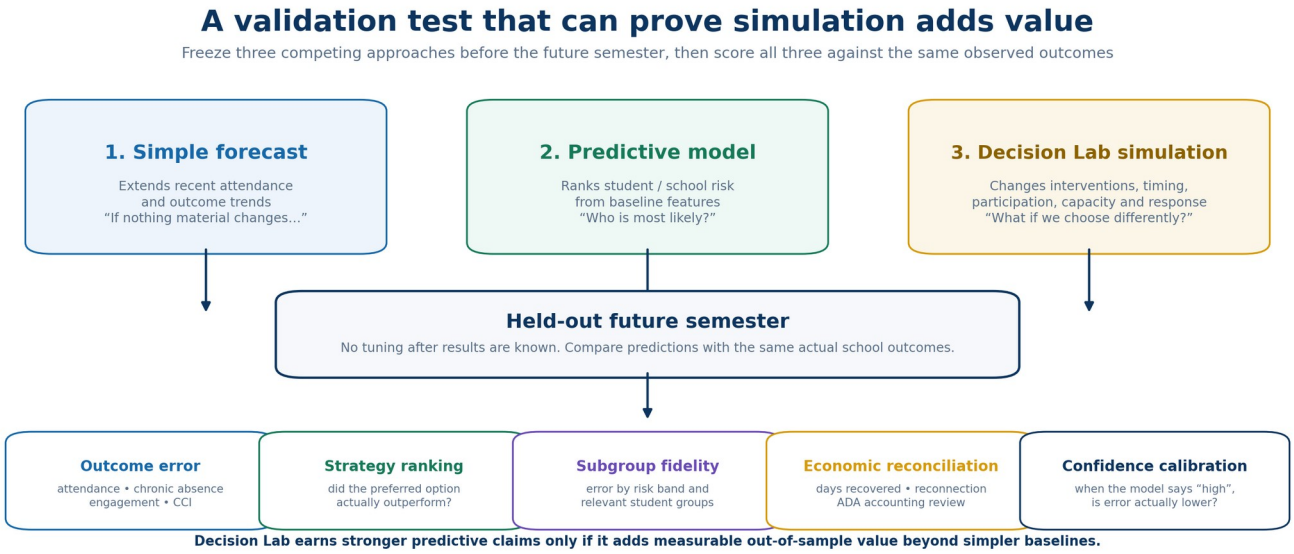

Figure 5. Proposed benchmark design. Forecast, predictive model, and Decision Lab simulation are frozen before a held-out future semester and scored against the same observed outcomes.

6.7 Scenario-level confidence

Average model performance is not enough: a system can perform well overall and still be unreliable for a particular school, subgroup, or strategy. Recent industry work at Simile explicitly treats question-level confidence as a separate model capability, estimating expected error for an individual simulation output [13]. Decision Lab should adopt the same general discipline without claiming the same model.

The first implementation can be evidence-based and categorical: High confidence when the school has fresh local data, strong comparable intervention evidence, stable sensitivity results, and adequate subgroup support; Moderate when evidence is strong but local response data are limited; Experimental when key response assumptions come mainly from educator estimates or synthetic calibration. After enough prospective runs accumulate, these labels should be replaced or supplemented by empirically calibrated error estimates.

6.8 Transportability and subgroup fidelity

A model that works on average can still fail for a particular school or student group. Version II therefore treats transportability as an explicit validation question: when can a response function learned at one school be reused elsewhere, and when is school-specific recalibration necessary? Subgroup error should be monitored by baseline risk band and other relevant student groups, subject to privacy, sample-size, and fairness safeguards. SocioBench findings that general-purpose LLM behavior simulations can show low accuracy and subgroup variation reinforce the need for domain-specific validation rather than assumed generalization [14].

7. How the model is intended to learn

Decision Lab's long-term advantage is intended to come from a data loop that ordinary public datasets cannot provide. Public data can describe enrollment, attendance, chronic absence, college/career measures, and other outcomes. They usually cannot tell a district exactly which supports a student saw, whether the student started them, whether the student followed through, when participation stopped, or how response changed over time.

The Campus Engagement System is designed to create those school-authorized signals as a by-product of services students actually use: planner activity, peer acknowledgement and seconding, homework routines, participation, family visibility, intervention exposure, and response. These signals are intended to support the next generation of model calibration, not to become a behavioral-surveillance system or a hidden individual score.

The Educator Research Questionnaire provides a second calibration stream. Educator judgments can be collected in a structured, versioned form and compared with current model assumptions. Cleanly mapped responses may update start, completion, continuation, and retention/fade assumptions; other answers should remain research inputs until a defensible mapping is established.

7.1 Evidence hierarchy for parameter updates

Highest	Prospective local implementation with well-defined exposure and outcome measurement
High	Replicated intervention research in comparable high-school populations
Moderate	Longitudinal observational relationships + local benchmark fit
Exploratory	Structured educator estimates and pilot feedback
Lowest	Unreviewed judgment or convenience assumption

The model should record the source, date, version, uncertainty, and intended use of every material assumption. This follows the AQuA Book emphasis on data, assumption, decision, and assurance logs [10].

7.2 CES as an intervention-response dataset

The long-term data asset is not merely an engagement log. The higher-value dataset links a defined school action to the student's observed response and subsequent outcome: intervention → exposure → start → completion or stop → continuation or fade → attendance / academic / engagement outcome. Repeated over time, this creates a longitudinal intervention-response dataset that public administrative files generally do not contain.

This dataset can support several model improvements: local response curves, better estimates of participation and completion, detection of intervention fatigue or fade, subgroup-specific calibration, sequence effects, and eventually learned response models. Any such use must remain school-authorized, de-identified or appropriately governed, purpose-limited, and consistent with district privacy and student-data requirements.

7.3 Learned response models are a future layer, not a present claim

The 2025 SocSci210 work demonstrates that models trained on large experimental-response datasets can generalize better to unseen studies and conditions [8]. That result supports a research direction for Decision Lab, not an immediate product claim. A future version may replace selected hand-set response functions with statistical or machine-learning models trained on sufficiently large, permissioned intervention-response data. Those models should be evaluated out-of-sample against simpler baselines before deployment.

7.4 Refresh, drift, and versioned learning

Student populations, school policies, staffing, calendars, and participation norms change. A validated model can drift. Decision Lab should therefore record data freshness, parameter version, evidence source, last validation date, and observed drift for every major outcome family. A parameter update should never silently rewrite prior scenario results; the system should preserve the model version used for each saved scenario and report when a newer calibration materially changes the expected result.

8. What Decision Lab can and cannot claim today

Appropriate current claims

✓ Decision Lab can combine official school data, research anchors, user assumptions, and synthetic student states to compare scenarios.

✓ The experimental engine can track heterogeneous synthetic students across multiple semesters with response variance, attrition, stopping, continuation, and fade.

✓ The system can show how modeled outcomes change when leaders change assumptions or strategy mixes.

✓ The architecture is designed to compare modeled outcomes with actual implementation results and recalibrate over time.

✓ Version II defines a prospective benchmark in which the simulation can be tested against simpler forecast and predictive baselines on held-out school outcomes.

✓ The architecture identifies intervention exposure, dosage, timing, start, completion, stopping, continuation, fade, and peer/network context as separate model inputs where data are available.

Claims that require more evidence

— That a modeled intervention will cause the projected attendance, engagement, CCI, or ADA change at a particular school.

— That the current simulation is more accurate than competing products or professional judgment in live district decisions.

— That synthetic student behavior is a validated digital twin of individual students.

— That an engagement signal proves learning, attendance, or academic improvement.

— That modeled ADA value equals realized district funding in the same period.

— That the current system has a calibrated question-level confidence score. Confidence estimation is a roadmap capability until enough held-out validation runs exist.

— That current peer/network effects are validated causal effects. The network extension is a modeling target that must be tested prospectively.

— That CES data are already sufficient to train a general student-behavior foundation model. The present objective is a bounded high-school intervention-response model.

8.1 Intended use

Decision Lab is presently best used as a structured decision rehearsal: define a baseline, make assumptions explicit, compare interventions, identify sensitive parameters, choose a practical plan, and create a measurement plan before implementation. It is not a substitute for professional judgment, statutory requirements, district financial review, or a causal program evaluation.

9. Research and validation roadmap

Phase 1 - Baseline integrity	Automated school/district data refresh; lineage checks; benchmark reconciliation; public-data documentation.
Phase 2 - Benchmark baselines	Build and freeze simple trend forecasts and predictive-risk comparators for every prospective pilot.
Phase 3 - CES instrumentation	Capture authorized intervention exposure, timing, dosage, participation, completion, stopping, continuation, peer context, and response.
Phase 4 - Structured local experiments	Use randomized encouragement, phased rollout, or other approved designs to estimate selected response functions without withholding required support.
Phase 5 - Prospective pilots	Pre-register scenarios; run matched simulations; hold out future semesters; measure actual implementation and outcomes; quantify error by outcome and subgroup.
Phase 6 - Confidence layer	Calibrate scenario-level confidence against observed error, model drift, evidence strength, and sensitivity stability.
Phase 7 - Replication and transportability	Repeat across multiple high schools and districts; test where response functions transfer and where local recalibration is required.
Phase 8 - Learned response + multi-agent extension	Where sample size and governance permit, train response models on intervention-response data and test network-aware student interactions against simpler baselines.

9.1 Proposed primary validation metrics

Calibration error	Modeled - observed aggregate outcome at fixed horizons
Rank-order utility	Whether strategy ordering matches observed performance
Sensitivity stability	Whether the preferred plan survives plausible parameter changes
Subgroup error	Calibration by baseline risk band and relevant groups
Implementation fidelity	Planned vs. delivered reach, start, completion, and retention
Economic reconciliation	Modeled vs. actual days/value plus district finance review
Forecast lift	Out-of-sample error improvement versus a simple trend forecast
Predictive-baseline lift	Out-of-sample error improvement versus a predictive/risk model
Confidence calibration	Higher-confidence outputs should realize lower observed error
Transportability	Performance at a new school or subgroup without local refitting
Sequence / dosage fidelity	Match between modeled and observed timing / dose-response

References and data sources

[1] California Department of Education. Chronic Absenteeism Downloadable Data Files, including 2024-25 state, district, school, and student-group files. Source

[2] California Department of Education. College/Career Indicator Downloadable Data Files, 2024-25. Source

[3] California Department of Education. 2024-25 End-of-Year Reports; chronic absenteeism and average days absent. Source

[4] National Center for Education Statistics. Education Longitudinal Study of 2002 (ELS:2002). Source

[5] National Center for Education Statistics. High School Longitudinal Study of 2009 (HSLS:09). Source

[6] Park, J. S., et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. UIST 2023. Source

[7] Park, J. S., et al. (2024). Generative Agent Simulations of 1,000 People. Source

[8] Kolluri, A., Wu, S., Park, J. S., & Bernstein, M. S. (2025). Finetuning LLMs for Human Behavior Prediction in Social Science Experiments. EMNLP 2025. Source

[9] Institute of Education Sciences. Getting Students on Track for Graduation: Impacts of the Early Warning Intervention and Monitoring System After One Year. Source

[10] UK Government Analysis Function. The AQuA Book: Guidance on Producing Quality Analysis (2025 edition). Source

[11] Park, Joon Sung. “Simulation: the new Scaling Law - Joon Sung Park, Simile AI.” Latent Space, August 21, 2026. Source

[12] Bernstein, Michael. “Building the What-If Machine.” Simile, September 2, 2026. Source

[13] Wesel, Andrew; Chen, Sarah; Liang, Percy. “Building confidence in Simile.” Simile, August 25, 2026. Source

[14] Wang, Jia; Zhao, Ziyu; Ni, Tingjuntao; Wei, Zhongyu. “SocioBench: Modeling Human Behavior in Sociological Surveys with Large Language Models.” EMNLP 2025. Source

Document status

Version: v0.2 (Version II). This document describes the current research architecture and validation roadmap as of September 2026. Version II adds behavior-grounded intervention-response data, experimental learning, peer/network effects, explicit forecast benchmarks, held-out validation, and a scenario-level confidence roadmap. It should be updated when model versions, data sources, validation evidence, or intended-use claims materially change.

Decision Support Labs is not affiliated with the California Department of Education, NCES, IES, Stanford University, Simile, Latent Space, the Association for Computational Linguistics, or the UK Government Analysis Function. Names and sources are referenced for data provenance, research context, and industry-methods comparison. Simile practices cited here are design inspiration and do not constitute validation of Decision Lab.