

OPERATING CASE STUDY / OPERATIONAL EXCELLENCE

Network Quality Governance

When quality scores outrun evidence

An anonymized project case study of a delivery network's quality-review mechanism: how metric definitions, action verification and source precision shape operating decisions.

PROJECT CONTEXT / DISTRIBUTED DELIVERY OPERATIONS

BLUF / BOTTOM LINE UP FRONT

Keep the operating cadence.
Validate evidence before assigning impact.

RECALCULATED FINDING

Five of six final-period measures move in an adverse direction. The only tentative sign of a process change is sensitive to source rounding.

CONTRIBUTION IN THE SOURCE RECORD

Analytics, weekly tracking and the review narrative.
Registry-attested role.

EVIDENCE STATUS

Descriptive operating case. Historical causal impact remains unestablished.

BLUF: analytical inference. Source contribution: historical record. Client and internal identities anonymized.

EXECUTIVE DECISION / LS-001

When quality scores outrun evidence

[AI] A distributed-operations case on measurement, intervention and decision control.

CONTEXT

[HR] A distributed delivery operator concentrated a sprawling quality agenda into a small set of operational inputs, a ranked focus cohort and recurring action reviews. The source registry assigns the analytics lead responsibility for the read, tracker and discussion narrative; a separate sponsor led the program.

OBJECTIVE

[AI] Determine what the mechanism supports as a management decision.

APPROACH

[RC] Reconcile six weekly observations for each of six inputs, calculate provisional process limits, examine failure-rate complements and test source-version and rounding sensitivity.

FINDINGS

[RC] Group A's sort-failure frequency rises 19.05% from T1 to T6; Group B's falls 16.95%. Five of six final-period movements are adverse in direction, but only one triggers a provisional process signal. That signal disappears under plausible rounding reconstructions. The directional story exceeds the evidence for a changed process.

DECISION VALUE

[AI] Authorize a targeted check of whether Group A's apparent disruption survives unrounded data. Hold any claim of program-caused customer improvement, cohort graduation based solely on score movement, or financial benefit booking until the population, action delivery and mature outcome reconcile.

IMPLICATIONS

[PD] The sponsor assigns the measurement owner and analytics lead to recover unrounded values and reconcile eligibility. Release diagnostic action with valid measurement and execution proof. Release causal or financial claims only after an identified mature-outcome effect and a validated incremental-cost bridge.

THE DECISION IN ONE SENTENCE

[AI] Keep the operating cadence, but replace "the number moved" with a traceable test of what changed, for whom, and after which delivered action.

Labels: HR historical record (registry); RC recalculated; AI analytical inference; PD proposed design. Source S1; calculation C1. Descriptive case, not an identified impact evaluation.

EXHIBIT 1 / OPERATING LOGIC

A score can direct attention; it cannot certify impact

[HR] The governance innovation was a repeatable conversation connecting a selected input to a named owner and an action.

Layer	Historical mechanism [HR]	Decision implication [AI]
Attention	A weighted input slate ranked operating units for focused reviews.	Ranking economizes management attention but selects extremes.
Accountability	A first review diagnosed the issue and committed actions; a subsequent review challenged whether the number moved.	An action owner is observable; causal credit requires more evidence.
Execution	The registry describes task tracking, evidence requests and escalation.	Assigned, completed and verified must remain separate states.
Learning	The mechanism evolved; the input slate and supporting forums changed.	Freeze each version before comparing its results across time.

[HR] The registry records problems in classification, eligibility and operational interpretation. A missing-item label could represent a workflow or return-status issue rather than permanent loss. Apparent sort defects could originate in physical handling rather than the scan step emphasized by a headline metric.

[AI] This changes the intervention. A taxonomy repair, a staffing change and a physical handoff repair act through different pathways. Combining them under one “quality improvement” label obscures whether the team changed reality, observation or both.

MANAGEMENT CONSEQUENCE

[PD] Preserve the ranked queue as a triage aid. Require each action card to name the failure mode, eligible population, execution proof, expected lag and mature outcome before the sponsor treats it as an impact claim.

Source S1: project narrative, implementation and decision records. Mechanism is registry-reported; primary artifacts were unavailable for direct verification.

EXHIBIT 2 / NEAR-CEILING MEASUREMENT

The pass-rate view hides a larger failure-rate movement

[AI] Re-expressing the same observations changes what deserves investigation without creating a causal effect.

THE FAILURE TAIL TELLS A DIFFERENT STORY

T1 = 100 independently within each metric and cohort; T6 shown

Metric view	Group A	Group B
Failure index at T6	119.05	83.05
Change T5 to T6 (%)	38.89	28.95

Derived failure-frequency change: A +19.05% | B -16.95%

Same observations. Descriptive transformation. No pooled or causal estimate.

- [RC] With each failure series’ own T1 set to 100, Group A ends at 119.05 and Group B at 83.05. In the final interval alone, failure frequency rises 38.89% and 28.95%, respectively. These endpoint comparisons expose the small failure tail of a near-ceiling success metric.
- [AI] A near-ceiling success metric compresses changes in a small failure tail. The complement is a more legible diagnostic for that tail. The transformation adds no observations, removes no confounding and cannot establish whether customers experienced fewer mature failures.
- [PD] Publish only the indexed failure view: dual complementary indices would reveal the hidden raw baseline algebraically. Compare within-cohort trajectories only. Independent rebasing erases original level differences and does not yield a network average or a between-cohort effect.

Source S1, stable six-period panel; calculation C1. Failure = 100% minus pass rate, calculated privately before each series is indexed.

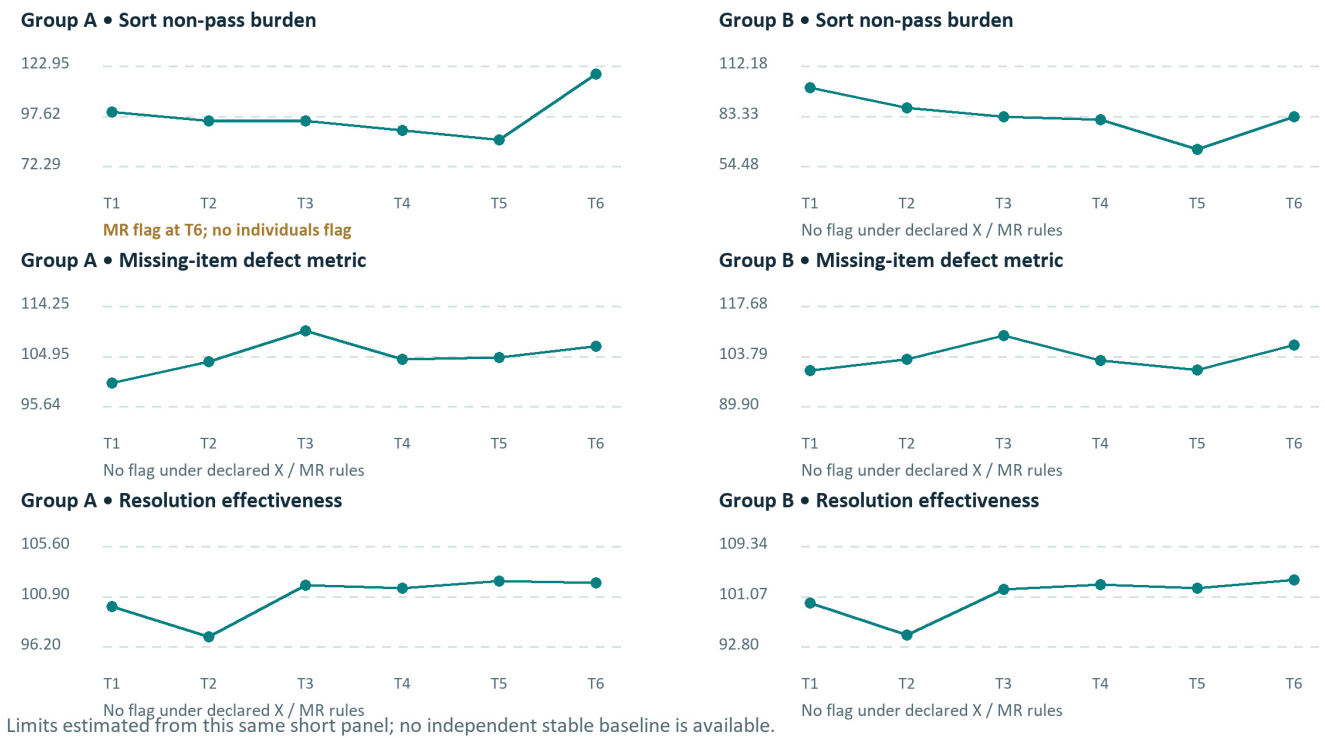
EXHIBIT 3 / PROCESS BEHAVIOR

One abrupt movement warrants validation, not a victory narrative

[PD] The uniform exploratory screen uses one observation outside three-sigma limits or one moving range above its limit.

SIX INPUT SERIES • PROVISIONAL INDIVIDUALS LIMITS

Each T1 = 100 | Independent vertical scales | Six observations per series



[RC] No individuals limit is breached. Group A sort alone crosses its moving-range limit at T6, at 1.0713 times the limit. Across 729 plausible rounding reconstructions, that ratio spans 0.8241–1.1903: both flag and no-flag results are possible.

[AI] “No signal detected” describes the other movements under this rule. The only flag is provisional and rounding-sensitive. The decision is to recover unrounded data and validate the event—not to declare a confirmed disruption from a short retrospective baseline.

Calculation C1: XmR, k=6 per series, MR span=2, d2=1.128, D4=3.267. Dashed lines: provisional individuals limits; separate MR flag shown.

EXHIBIT 4 / SOURCE RECONCILIATION

Conflicting values change the magnitude, not the validation decision

[PD] The reproducible weekly panel is the primary calculation basis; credible alternate cells remain visible.

T5 → T6 relative change [RC]	Primary panel	Alternate cells
Group A missing-item defect metric	+2.018%	+1.856%
Group A resolution effectiveness	−0.169%	−0.375%
Group B resolution effectiveness	+1.364%	+1.147%
Group A sort MR flag	Flag at T6	Unchanged

[AI] No alternate flips these directions or resolves the link between the action package and mature customer outcomes. A small version disagreement is not permission to select the more favorable value. The figures above are relative changes, not percentage-point differences.

[PD] Selection protocol: use the explicit same-definition weekly panel for reproducibility; retain each contemporaneous alternative; test all affected calculations; seek an authoritative event-level extract before certifying a final reported value. The current choice is a transparent analytical basis, not a finding that this source is authoritative.

[HR] The registry also records conflicting goals and unit-cost assertions, a later target on a different planning horizon and changes in the metric slate. [AI] A later target is not another contemporary observation. A revised score with undisclosed normalization cannot be chained onto the founding composite.

DO NOT IMPORT THE SOURCE’S CONCLUSION

[AI] A narrative attribution based on a standardized regression coefficient cannot be converted into raw defect reduction without scale information, a validated model and an identification argument. This case does not carry that conversion forward.

Source S1, explicit weekly panel and alternate-cell register; calculation C2. Exact private cell references and raw conflicting values are in the private annex.

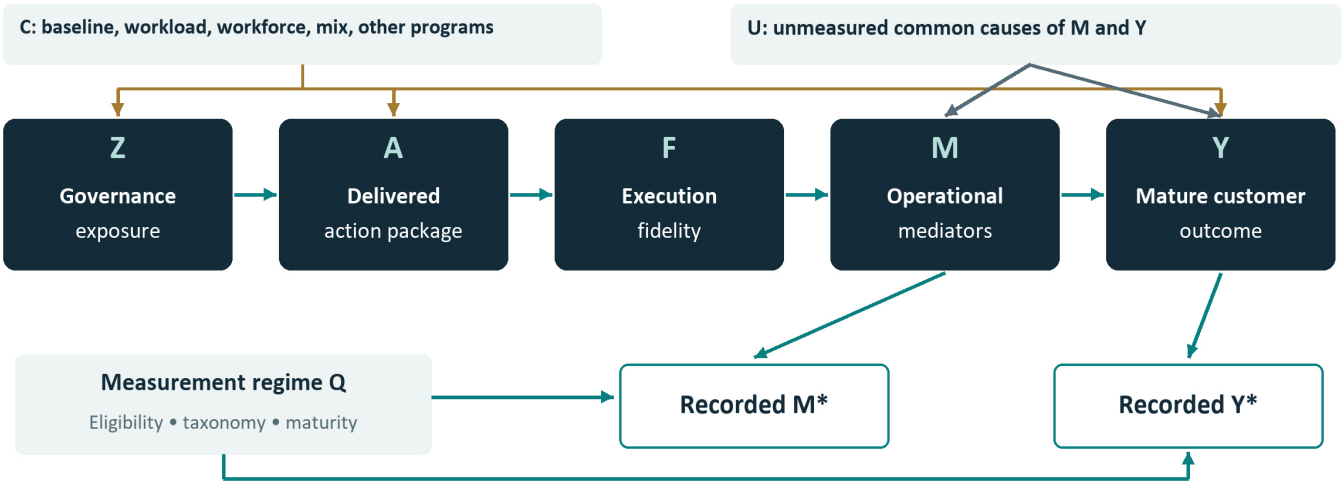
EXHIBIT 5 / CAUSAL ARCHITECTURE

The missing bridge is delivered action to mature customer outcome

[PD] The Validated Action-to-Outcome Chain separates governance exposure from the treatment customers could experience.

VALIDATED ACTION-TO-OUTCOME CHAIN

Proposed causal structure; arrows are assumptions requiring tests



Reduced graph: Q can change recorded measures without changing the physical process.

[PD] This reduced graph formalizes assumptions, not estimated effects. M aggregates the sort, resolution and dispatch mediators expanded in the technical ledger; additional direct paths appear there. C includes baseline performance, volume and mix, staffing, node type, business line, seasonality and concurrent programs. Q denotes the measurement regime.

[AI] The registry supports the existence of the governance and analytical work. It does not furnish the cluster-level assignment, action-fidelity and mature customer panel needed to identify the causal bridge. Adjusting for the mediator would change a total-effect estimand; conditioning on observed completion alone would select implementers.

RELEASE CRITERION

[PD] Pair each eligible unit’s assignment date with a versioned action package, execution evidence and an outcome observed after its full maturity window. Keep the governance-policy effect distinct from the effect of the delivered action.

Proposed DAG and framework: Luna Sol. Causal interpretation follows Hernán & Robins; historical mechanism basis S1. See technical edge and estimand ledger.

EXHIBIT 6 / ECONOMICS

Test detectability against break-even before approving rollout

[PD] Economic thresholds and evaluation precision must use the same population, horizon and common scale.

NET VALUE UNDER DECLARED PLANNING ASSUMPTIONS

Untreated loss = 100 | Cost = 2 | Non-overlapping eligible share = 0.75

Mature reduction	50% avoidable	75% avoidable	100% avoidable
5%	-0.12	+0.81	+1.75
10%	+1.75	+3.62	+5.50
15%	+3.62	+6.44	+9.25

Values are normalized net-benefit units, not currency or realized project savings.

If the avoidable share and eligible share are both 50%, break-even rises to 8.00%.

[PD] Normalize untreated eligible-period customer loss to 100 value units and incremental program cost to 2. With avoidable-cost share 0.75 and non-overlapping eligible share 0.75, a 3.56% mature outcome reduction breaks even. These are declared planning assumptions, not observed cost or benefit.

[PD] The heat map varies the mature relative reduction and avoidable-cost share while holding the non-overlapping eligible share at 0.75. A positive cell means a proposed economic scenario clears incremental cost; it does not establish the intervention can produce that reduction.

[AI] A review that cannot reliably detect the economically relevant reduction cannot settle the rollout decision. The technical power model therefore tests candidate rollout schedules against this same threshold and preserves the no-go result when precision is insufficient.

Calculation C4: net value = 100 × relative reduction × avoidable share × non-overlapping share – 2. All economic parameters are proposed design.

EXHIBIT 7 / DECISION INFORMATION

Validation has value only when it changes the preferred action

[AI] A correction can protect an allocation before any customer-outcome benefit is demonstrated.

[HR] The registry records analytical corrections, maturity warnings and population-definition issues. It does not establish a transaction-level financial consequence for those decisions. [PD] A separate information model therefore prices a hypothetical validation decision rather than assigning historical savings.

Model element [PD]	Decision contract
State	The decision premise is valid or invalid.
Actions	Authorize a proposed resource commitment or defer it.
Information	A validation check returns pass/fail with stated sensitivity and specificity.
Rule	Authorize only when posterior expected decision payoff exceeds deferral.
Value	Expected best payoff after the check minus best payoff before it; then subtract check cost.

[AI] Information that leaves the chosen action unchanged can have zero expected value. Greater diagnostic accuracy does not automatically justify a costly review. The low/base/high cases in the technical model expose this trade-off and satisfy the perfect-information bound.

ACCOUNTING BOUNDARY

[PD] Report decision-information value separately from program outcome value. The examples use abstract decision-payoff units, concern authorization error and are not added to the operating loss index in Exhibit 6.

Proposed decision model C5; method: Strong et al. (2015). No measured historical information value or realized savings is asserted.

TECHNICAL FOUNDATION · CAUSAL MODEL

A measured input is a mediator candidate, not proof of a customer benefit

[AI] The causal model separates review assignment, implementation, execution fidelity, operational mediators, customer outcomes and measurement artifacts.

[AI] Let Z denote review assignment; A action-package implementation; F execution fidelity; M operational mediators; and Y mature customer defects. The overview DAG is $Z \rightarrow A \rightarrow F \rightarrow M \rightarrow Y$. Expand M as {I, R} $\rightarrow$ D: I denotes sort success, R resolution effectiveness and D dispatch defects. Additional $I \rightarrow Y$, $R \rightarrow Y$ and $A \rightarrow Y$ paths are hypotheses omitted from the overview.

Directed path	Interpretation and evidence boundary
$Z \rightarrow A \rightarrow F$	[HR] The registry describes assigned actions. [AI] Assignment, implementation and verified execution remain distinct.
$F \rightarrow I, R, D$	[AI] Executed staffing, flow and exception closure could change operational performance. Weekly association does not identify these edges.
$I, R, D \rightarrow Y$	[AI] These are the intended customer-benefit paths. No matched mature-outcome panel identifies them here.
$C \rightarrow A, F, M, Y$	[AI] Demand, weather, staffing, package mix, node type, calendar effects and concurrent programs are confounders.
$Q \rightarrow I^*, R^*, D^*, Y^*$	[AI] Taxonomy, denominator and extraction changes affect reported measures; starred variables are observations.
$B^* \rightarrow G \rightarrow Z$	[AI] A noisy baseline composite selects the focus cohort G, creating selection and mean-reversion risk.

[AI] Unmeasured factors U may affect adoption, mediators and outcomes. A model residual combines omitted paths, model error and noise; it cannot establish named causes. Do not adjust the total-effect analysis for post-assignment mediators or fidelity.

[PD] Register every effect claim against an identified edge or total path, a validated measure and a declared estimand. Otherwise retain it as a hypothesis.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION • ESTIMANDS

Separate the effect of the review offer from the effect of the executed action package

[PD] Define population, treatment, endpoint and maturity before estimation. An input change is not a customer-impact estimate.

[PD] Lock eligible sites and baseline definitions before rollout. Retain noncompliers and later graduates. Fix site weights from the pre-period; prespecify shipment-weighted sensitivity.

Quantity	Definition and interpretation
Review-offer effect	[PD] $ITT(H) = E[Y_i(1,H) - Y_i(0,H) \mid \text{eligible}]$. Treatment is randomized access to the review-and-action protocol; outcome is mature customer-defect incidence over horizon H.
Action-policy effect	[PD] $E[Y_i(a^*,H) - Y_i(a_o,H) \mid \text{eligible}]$. a^* specifies staffing protection and exception closure; a_o is current practice. This differs from an effect among voluntary adopters.
Mediator effects	[PD] Estimate separate ITTs for I, R and D over aligned windows. Operational responses do not replace the customer endpoint.
Heterogeneity	[PD] Estimate treatment-by-baseline-stratum interactions; comparing subgroup significance does not test differences.

[PD] Fix H before assignment. Apply one closure lag L, based on historical reporting completeness, to every shipment. Freeze the snapshot after final eligibility plus L; test a longer-lag sensitivity.

[AI] Aggregate historical observations do not identify these estimands. Voluntary adoption is confounded; governance also has pathways beyond the specified action package, so an instrumental-variable exclusion restriction is unestablished.

[PD] The financial planning example assumes fixed mean loss per defect and unchanged severity mix, so incidence and loss share one baseline-100 proportional scale. If these assumptions fail, preregister a fixed-cost-weighted burden estimand and recompute power and economics.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION • MEASUREMENT VALIDITY

A stable label is insufficient when the denominator or classification changes

[AI] Valid measurement precedes evaluation. The registry reports basis and population-mapping changes.

Construct	Required validity checks
First-pass sort success	[PD] Define eligible processing events, first-pass timing and treatment of rework. Audit whether rescans reset eligibility. Confirm the same event definition in both cohorts. Report the failure complement to expose ceiling effects.
Resolution effectiveness	[PD] Define an eligible problem, successful resolution and observation lag. Reconcile repeat cases and unresolved cases. Check whether adoption or staffing changes alter which problems enter the denominator.
Dispatch-availability defects	[PD] Distinguish physical absence, scan failure, misclassification and later return. Use a fixed dispatch denominator. Test taxonomy agreement across nodes and reporting periods.
Ranking composite	[PD] Specify transformation, direction, missingness, weights and eligibility. Test rank sensitivity to each choice. Raw percentages and defect rates cannot be combined simply because weights sum to one.
Mature customer defects	[PD] Define unique shipment-level outcomes, category overlap, reporting lag and financial eligibility. Separate actual service failure from compensation-policy and reporting changes.

[HR] Source node types were labeled as business lines. [PD] Retain neutral cohort A/B labels until a versioned crosswalk verifies the mapping; business-line heterogeneity is currently unidentified.

[PD] For taxonomy reliability, draw a stratified random event sample covering cohorts, time windows and defect categories. Two trained reviewers independently code the same masked records using a frozen handbook; adjudicate only after independent coding. Choose sample size to attain a prespecified confidence-interval width for agreement. Report raw agreement, category prevalence, confusion matrices and Cohen’s κ with uncertainty; κ alone can mislead when categories are rare.

[PD] Revalidate after eligibility, taxonomy or extraction changes. Preserve an overlap sample to bridge measurement regimes instead of splicing series silently.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION · PERSONAL CONTRIBUTION

Attribute the analytical read to its owner without claiming the entire mechanism

[HR] The registry describes an analytics lead responsible for tracker production, call framing, measurement corrections and synthesis. It assigns charter ownership, forum leadership and field execution to other roles.

Contribution claim	Process-tracing test and present grade
Produced the analytical read	[HR] The registry quotes creator metadata and owner headers. [AI] Hoop-test support is present at registry level; direct-original verification remains open. Authorship would establish production, not customer impact.
Identified a goal-basis inconsistency	[HR] The registry describes a correction and arithmetic back-solve. [AI] Hoop support for an analytical trace; personal attribution remains registry-attested. No verified decision-adoption smoking gun is available.
Changed leadership decisions	[AI] A delivered analytical artifact is insufficient. Require a dated decision record linking the finding to a changed or deferred action. Current support is incomplete.
Delivered network outcomes	[AI] Not established. Site execution and competing programs create alternative explanations; no doubly-decisive test or identified causal outcome estimate is available.

[AI] A hoop test asks whether evidence required by the proposed contribution is present; passing it does not eliminate rival explanations. A smoking-gun trace is highly specific support for a mechanism but need not be necessary. A doubly-decisive test both supports the explanation and rules out plausible rivals. These labels describe inferential strength, not a points-based personal performance score.

[PD] Separate artifact authorship, analytical judgment, decision adoption, field execution and outcome causation. For each claim, record the source, rival explanation, test and verification gap. Metadata, an owner header and a summary copied from that file share one evidentiary lineage; they are not independent confirmations.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION · IMPLEMENTATION FIDELITY

A review can be delivered successfully while its action package is never executed

[AI] Meeting occurrence, tracker publication and action assignment cannot establish that operational behavior changed.

[PD] Define the action package before rollout: behavior, authority, eligibility, timing, minimum dose, evidence and exceptions. Keep the specification stable within a wave; version changes create a new intervention state.

Delivery layer	Auditable evidence and inference
Offer / exposure	[PD] Record assignment date, review access and receipt of the decision brief. Retain every assigned site in the ITT denominator.
Acceptance	[PD] Capture named role acceptance, resource authorization and an implementation start. These establish readiness, not execution.
Action execution	[PD] Measure eligible opportunities and verified completed actions through timestamps and independently sampled operational records. Self-attestation alone is insufficient.
Operational response	[PD] Observe aligned input measures after a prespecified response lag. Compare with the assigned control policy; separate execution failure from a weak causal mechanism.
Customer response	[PD] Observe mature outcomes after the action-to-service lag and reporting closure. Require guardrail checks on cost, safety and service displacement.

[AI] Four interpretations follow: low fidelity leaves the package untested; high fidelity with no input response challenges the immediate mechanism; input response without customer response challenges mediation or offsetting paths; customer response with adverse guardrails questions net decision value. These are diagnostic classifications, not post hoc justifications for a preferred result.

[PD] Report fidelity by randomized assignment and baseline strata. A secondary per-protocol analysis requires explicit adjustment for time-varying confounders of adherence and outcomes; it must not replace ITT. Document spillovers when staff, practices or capacity cross between sites. Treatment contamination can shrink the observed contrast even when the underlying operating action is effective.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION · SELECTION AND MODERATORS

Worst-cohort selection can produce apparent recovery without an intervention effect

[AI] Extreme performance combines persistent weakness with temporary noise. Re-measurement can remove the noise without operational improvement.

[AI] Under classical reliability assumptions, $E[X_2 \mid X_1] = \mu + \rho(X_1 - \mu)$; expected reversion is $(1 - \rho)(\mu - X_1)$. This sensitivity model is not estimated here: untreated repeated site measures are missing. Extreme selection creates risk, not guaranteed recovery.

[PD] Quantify that risk using untreated historical cohorts formed by the same frozen selection rule. Estimate selection-period reliability, repeat the ranking across alternative baseline windows and simulate expected recovery absent treatment. Preserve the original selected cohort after graduation; removing recovered sites changes the estimand and introduces survivor bias.

Moderator	Testable reason the response may differ
Node and workflow design	[AI] Different process steps can alter whether an action reaches the constrained stage. Validate the node crosswalk before classifying subgroups.
Baseline capacity	[AI] Protecting problem-resolution time may work where capacity can be reallocated; a hard staffing constraint may require a different treatment.
Defect composition	[AI] A debrief action can affect misclassification and closure behavior; it need not reduce physical loss or upstream sort failure.
Shock exposure	[AI] Weather, volume surges and package mix may change both feasible action dose and outcome risk. Prespecify interactions and aligned controls.

[PD] Compare focus and non-focus units only after matching definitions, maturity windows and calendar periods. Where selection is deterministic, compare units near the eligibility threshold as a supplementary design if manipulation and continuity checks pass. A network average is not an untreated control when the same actions diffuse across the network.

[AI] Aggregate source groups cannot identify subgroup effects or establish that a narrower action list works better.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION • PROSPECTIVE EVALUATION

Randomize the rollout order before using the scorecard to claim causal credit

[PD] Randomized phased rollout can preserve eventual access while separating the protocol’s effect from calendar trends and cohort selection.

[PD] Randomize eligible clusters within prespecified node and baseline-risk strata to rollout sequences. Use sites as clusters only if staffing, management and treatment spillovers are sufficiently contained; otherwise randomize the shared operating unit. Fix the action package and collect an untreated pre-period before the first wave.

Design element	Prespecified requirement
Primary contrast	[PD] Estimate the review-offer ITT on the mature customer endpoint. Include calendar-period effects and account for clustering and repeated observations.
Adoption timing	[PD] Record assigned and actual starts separately. Prevent anticipation where possible and test for it. Calendar shocks require concurrent controls.
Event-time diagnostics	[PD] Inspect pre-treatment outcome paths, composition and missingness. In an observational rollout, assess conditional parallel trends and overlap; a nonsignificant pre-trend test does not establish validity.
Heterogeneous timing	[PD] For a justified staggered observational design, estimate cohort-by-time effects with not-yet-treated or never-treated controls; avoid a single two-way-fixed-effects coefficient that can mix heterogeneous effects.
Power and decision threshold	[PD] Simulate the actual rollout with cluster-size imbalance, serial correlation, intracluster correlation, outcome maturity and attrition. Compare the MDE with economic break-even before launch.

[PD] If the test cannot resolve break-even, increase independent clusters, improve baseline precision, extend the useful observation window, reduce measurement error or redesign the decision. Additional shipments inside already-large clusters may add little information. A positive but imprecise estimate is not a scale decision.

[AI] Hussey and Hughes inform stepped-wedge analysis; Callaway and Sant’Anna address heterogeneous staggered effects. Their identification assumptions require substantive evidence. This is a prospective protocol, not a completed evaluation.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION · CONTRIBUTION AND PROPOSITIONS

The Validated Action-to-Outcome Chain makes operating claims falsifiable

[AI] The proposed Validated Action-to-Outcome Chain links five gates: valid measurement, stable population, verified action, mature outcome and incremental economics. A failure at an earlier gate limits what later evidence can establish.

[AI] This case is consistent with Shewhart and Deming's concern that reacting to routine variation can destabilize a process. Campbell's analysis of indicators explains why accountability can change recorded classifications as well as performance. The metric-basis conflicts illustrate the need to validate the measurement chain before praising the operating chain; they do not independently establish intentional gaming. [AI] Goodhart's warning about relationships changing under control is a conceptual analogue. This record neither tests the monetary proposition nor proves deliberate gaming.

[AI] Ittner and Larcker's work on nonfinancial measurement motivates testing the assumed links between leading indicators and value. Tucker and Edmondson's work on learning from operational problems motivates tracing whether recurring exceptions generate organizational correction rather than repeated workaround. Composite-indicator guidance from the OECD and Joint Research Centre motivates sensitivity to scaling, weighting and aggregation.

[PD] P1 · Measurement gate. A validation gate reduces incorrect escalation relative to usual review. Falsify if independently adjudicated decision-error rates do not fall under randomized assignment.

[PD] P2 · Execution gate. Randomized delivery assurance increases action fidelity and improves the prespecified mediator. Falsify if assignment raises verified fidelity yet its ITT excludes the required mediator response. Do not condition this test on observed completers.

[PD] P3 · Outcome gate. Mediator response translates into mature customer benefit within the specified lag and population. Falsify if identified mediator changes coexist with a precise exclusion of the required customer benefit.

[PD] P4 · Decision-value gate. Validation costs less than the expected decision loss it prevents. Falsify if audited avoided-error value remains below incremental validation cost over the registered horizon.

[AI] The contribution is a testable synthesis for this operating setting, not a claim of a newly established universal law. The case extends the practical discussion by showing that directionally stable executive headlines can coexist with irreconcilable effect magnitudes and unsupported causal paths. Formal novelty and external validity require further study.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION · LIMITATIONS

The evidence supports a governance diagnosis, not an identified historical impact estimate

[AI] The limitations below define the inferential ceiling of this case. They are not resolved by a more polished narrative, a larger modeled benefit or a favorable input trajectory.

Evidence lineage

[HR] Historical facts are drawn from the supplied registry and its transcriptions. Referenced original artifacts were not available for direct reopening in this review. [AI] Repeated summaries can share one source lineage; their agreement does not establish independent corroboration. Personal-contribution claims remain registry-attested.

Measurement and statistical identification

[HR] The record contains competing value versions, a changed metric basis, uncertain node-to-business mapping and incomplete maturity and denominator information. [AI] Six aggregate observations provide weak process-limit estimates and no untreated counterfactual. Normal-looking variation cannot establish no effect; an out-of-limit point cannot identify the cause. Rate-specific limits and measurement-invariance tests require counts and event data absent here.

Selection and treatment

[AI] Extreme-cohort selection, graduation, concurrent programs, unequal action fidelity and operating shocks can explain part or all of observed movement. The registry does not provide a stable site panel or independent control comparison. Neither a mechanism's existence nor artifact publication proves that field actions occurred or persisted.

Economics and decision value

[HR] Unit-cost assertions conflict, and incremental program cost, avoidable shares and realized customer impacts are not established. [AI] Scenario economics and information value are proposed decision aids; they are not historical benefits. Their probabilities, costs and effect sizes require elicitation or estimation before investment use.

Transferability and confidentiality

[AI] Source-group differences cannot be generalized to business lines without a verified crosswalk. Proposed methods require adaptation to local treatment constraints and outcome lags. Public indexing preserves within-series relative change but conceals absolute scale and prevents some cross-series comparisons. The private replication annex retains the mapping, raw transcription and calculations for authorized verification; public readers can audit logic but cannot independently authenticate the inaccessible originals.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

TECHNICAL FOUNDATION • REPRODUCIBILITY CONTRACT

Every published number needs a source, transformation and permitted interpretation

[PD] Separate private replication from public indexing. Rebuilding calculations does not authenticate historical sources.

Artifact	Minimum contract
Source manifest	[PD] Record source identifier, immutable hash, version, retrieval status, worksheet/cell locator and evidence lineage. Mark transcribed records separately from reopened originals.
Metric dictionary	[PD] Define construct, unit, numerator, denominator, eligible population, event time, reporting time, maturity lag, direction and measurement version.
Analysis ledger	[PD] Store raw input, selected version, calculation, rounding rule, public index base, claim class, sensitivity alternative and intended use for each exhibit value.
Treatment panel	[PD] Record locked eligibility, cluster, assigned wave, actual start, action version, verified dose, spillovers, exposure and mature outcomes without overwriting history.
Executable replication	[PD] Provide deterministic scripts, dependency versions, machine-readable outputs and automated reconciliation checks. Assert every public value against its calculation ledger.
Conflict log	[PD] Preserve rejected and alternative values with reasons. Re-run headline conclusions under each coherent version; do not combine incompatible versions to manufacture precision.

[RC] A within-series index is $100 \times x_t/x_o$. Compute a sort-failure index from $100 \times (1 - p_t)/(1 - p_o)$, where p is expressed as a proportion. Indexing by a positive fixed constant preserves relative movements and individual-chart signal classification; it does not preserve raw effect units across differently scaled series. Keep the index base and transformation fixed in all sensitivity runs.

[PD] Use [HR] for historical-record statements, [RC] for recalculation, [AI] for analytical inference and [PD] for proposed design. Treat references as support for methods, not evidence that this operator achieved those methods’ results. A final machine check should reject unlabeled scenario assumptions, unsupported attribution, exposed identifiers, unreconciled arithmetic and claims that exceed the declared estimand.

Labels: [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source basis S1; see private annex.

CALCULATION C1 / VARIATION & ROUNDING

The only detected signal is sensitive to source precision

[RC] Tentative XmR results use the transcribed six-period series, transformed to each series’ own T1 = 100.

Series [RC]	Mean	X lower	X upper	Mean MR	MR upper
A / Sort non-pass	97.62	72.29	122.95	9.52	31.11
A / Missing-item metric	104.95	95.64	114.25	3.50	11.43
A / Resolution	100.90	96.20	105.60	1.77	5.77
B / Sort non-pass	83.33	54.48	112.18	10.85	35.44
B / Missing-item metric	103.79	89.90	117.68	5.22	17.07
B / Resolution	101.07	92.80	109.34	3.11	10.16

[RC] Let x be the indexed value and $MR(t) = |x(t) - x(t-1)|$. Estimate sigma as $mean(MR)/1.128$. Individuals limits are $mean(x) \pm 3 \text{ sigma}$; the moving-range upper limit is $3.267 \text{ mean}(MR)$. Six values yield five adjacent ranges. No missing periods are interpolated and no supplemental run rules are applied.

[RC] All individuals observations lie within these limits; only A-sort’s last MR exceeds its limit. Deterministic rounding sensitivity shifts each underlying source value to just inside its lower bin edge, its printed value and just inside its upper edge: $3^6 = 729$ reconstructions. Both signal and no-signal results occur. The grid is not a probability distribution.

[AI] Source precision is material to the decision. Recovering unrounded data takes precedence over an explanation of a supposedly confirmed special cause. The affine complement and indexing preserve the X/MR flag sets, while protecting the hidden success-rate baseline.

[PD] Laney p' remains unestimated: event counts and eligible denominators are unavailable. Obtain them first, assess extra-binomial dispersion and rational subgroups, and distinguish binary nonconformance from multiple defects per opportunity, which require a different count model.

C1: deterministic reproduction; methods NIST, Wheeler and Laney. Short, retrospective panel; no claim of process stability or causal effect.

CALCULATION C3 / SELECTION SENSITIVITY

Extreme selection can create a turnaround without treatment

[PD] A normal-theory sensitivity quantifies the risk; it does not estimate this project’s regression to the mean.

[PD] Suppose untreated baseline and follow-up outcomes are bivariate normal, have equal means and variances, and correlate at rho. Higher is worse. For an illustrative worst-15% selection, the threshold is 1.036 standard deviations above the mean and the selected mean is 1.554 standard deviations above it.

Untreated persistence rho [PD]	Expected untreated rebound (SD)	Selected excess reverting
0.25	1.166	75%
0.50	0.777	50%
0.75	0.389	25%

[RC] Under that proposed model, $E[Y1(0) \mid Y0] = \mu + \rho(Y0 - \mu)$. The selected-group rebound is $(1 - \rho) \times \phi(z_{0.85})/0.15$. The table is a deterministic consequence of the stated distribution and persistence assumptions; it contains no fitted project parameter.

[AI] A selected unit can leave an extreme rank even when no action works. A rotating focus group also changes composition. A network aggregate that includes focus units cannot serve as an independent untreated comparison, and a fraction of sites is not necessarily the same fraction of shipment volume.

[PD] Retain unit identifiers, selection thresholds, ranking dates, tie rules, fixed original membership, eligible volume and repeated untreated histories. Estimate persistence from those histories; perform placebo extreme selections before rollout; then compare mature outcomes for fixed cohorts. Do not subtract a hypothetical rebound from the observed series and call the remainder causal.

C3: illustrative bivariate-normal calculation; Barnett et al. (2005). SD units are proposed, not internal raw metrics.

CALCULATION C6 / EVALUATION PRECISION

The smaller rollout cannot resolve the economic decision

[PD] Design-specific Gaussian calculations compare detectability with a common-scale break-even effect.

Independent clusters [PD]	MDE (index points)	Power at break-even	Decision
24	7.34	27.40%	Redesign
48	5.19	48.42%	Redesign
108	3.46	82.13%	Conditional candidate

[PD] Assign clusters evenly to six randomized sequences across eight periods, switching at periods 2–7. The outcome is mature customer loss on one common hypothetical counterfactual = 100 scale. Assume an immediate constant effect, no attrition and independent cluster-period residual SD = 10. These are planning inputs, not historical estimates.

[PD] Use cluster and calendar fixed effects with known covariance $V(i) = 100 I + [100 \rho / (1 - \rho)] J$. GLS variance is the treatment diagonal of $(X' V^{-1} X)^{-1}$. Solve the two-sided normal power equation for 80% power at $\alpha = 0.05$. Break-even = 3.56 index points corresponds to a 3.56% reduction from the common 100 baseline.

[RC] ICC sensitivities of 0.10, 0.30 and 0.50 give the same MDE because cluster fixed effects absorb the constant random intercept while residual SD stays fixed. This is an algebraic property of this model, not evidence that correlation never matters. The smallest balanced count clearing the threshold is 108.

[PD] Do not authorize a field test from this number alone. Estimate serial covariance and lag from a pre-period, add unequal cluster size, attrition and finite-cluster inference, and simulate the frozen estimator before commitment. If 108 independent clusters are infeasible, improve precision or extend mature follow-up and recompute; retain “inconclusive” if break-even remains undetectable.

C6: known-covariance stepped-wedge planning model; Hussey & Hughes, Hemming et al. Complete design matrix and inverse-information checks in private code.

CALCULATION C4 / ECONOMIC CONTRACT

Value follows the mature estimand and a disjoint cost bridge

[PD] The economic model explicitly prices a customer outcome in the evaluation population.

[PD] Let N be eligible opportunities over the evaluation horizon; p_0 the untreated mature-defect probability; r the identified relative reduction; c gross loss per mature defect; a the avoidable share; q the share of eligible loss not already claimed by another program; and C incremental implementation and operating cost. Net value = $N p_0 r c a q - C$.

[PD] In the normalized example, $N p_0 c = 100$ and $C = 2$. Hence $rBE = 2/(100 a q)$. Outcome index and value index share the same population and horizon, with constant average loss per defect assumed. If the action changes severity or loss composition, estimate that separately and rebuild the bridge.

Avoidable share a [PD]	q = 50%	q = 75%	q = 100%
50%	8.00%	5.33%	4.00%
75%	5.33%	3.56%	2.67%
100%	4.00%	2.67%	2.00%

[RC] The table reports break-even mature relative reductions. At $a = q = 0.75$, proposed reductions of 5%, 10% and 15% yield normalized net values of 0.8125, 3.6250 and 6.4375. Doubling incremental cost doubles every break-even threshold. These values are scenarios, not confidence intervals.

[PD] q is a disjoint-population/financial-overlap adjustment, not a second causal discount. Set $q = 1$ if the causal estimand and ledger already exclude overlap. Set $a = 1$ if c is already net avoidable cost. Track release of capacity separately from cash removal; use contribution margin for revenue opportunities. Finance signs the cost basis and horizon before a benefit enters the ledger.

C4: all economic inputs are proposed normalized assumptions. No registry unit-cost assertion or realized savings figure is adopted.

CALCULATION C5 / BAYESIAN DECISION VALUE

A validation check can change an authorization decision

[PD] Information value is calculated from decisions and evidence accuracy, independently of an observed treatment effect.

[PD] Two states: a sound or unsound decision premise. Authorize yields +10 decision units if sound and -30 if unsound; hold yields 0. Authorize when posterior probability of soundness exceeds 0.75. The check costs 0.50 units, with no additional delay cost in this example.

Scenario [PD]	Prior sound	Sensitivity / specificity	EVSI	EVPI	Net check value
Low	0.80	0.70 / 0.70	1.80	6.00	1.30
Base	0.65	0.85 / 0.85	3.95	6.50	3.45
High	0.50	0.95 / 0.95	4.00	5.00	3.50

[RC] Base case: $P(\text{pass}) = 0.65 \times 0.85 + 0.35 \times 0.15 = 0.605$. $P(\text{sound} \mid \text{pass}) = 0.9132$; $P(\text{sound} \mid \text{fail}) = 0.2468$. Without information, hold is optimal. With information, authorize after pass and hold after fail. Expected payoff becomes 3.95 before the 0.50 check cost, producing net information value 3.45.

[RC] $EVSI = E_D[\max_{\text{action}} E(\text{payoff} \mid D)] - \max_{\text{action}} E(\text{payoff})$. The perfect-information value is an upper bound: every scenario satisfies $0 \leq EVSI \leq EVPI$. Low/base/high jointly vary priors and accuracy; they are declared decision scenarios rather than uncertainty intervals.

[AI] The registry establishes neither those probabilities nor those payoffs. The calculation shows what evidence a defensible decision-value estimate would need. It does not monetize a historical correction. Because the payoff concerns authorization error and uses a separate normalization, it is not added to operating benefits in C4.

C5: exact two-state Bayesian enumeration, all inputs proposed. Strong et al. (2015) supplies the formal EVSI framework; example is Luna Sol analysis.

Read the case as a reproducible decision analysis

[AI] Evidence classes distinguish what happened, what was recalculated and what remains a testable proposal.

Label	Meaning in this case
[HR] Historical record	The supplied project registry reports the fact. This is not a claim that its unavailable original attachments were reopened.
[RC] Recalculated	A deterministic transformation of supplied values or an explicitly stated proposed scenario. The surrounding label states which.
[AI] Analytical inference	An interpretation supported by the record and methods; not an identified causal estimate.
[PD] Proposed design	A future protocol, hypothetical parameter or falsifiable mechanism; not an implemented or observed result.

[AI] Action titles inherit their exhibit’s claim scope. Numbers in the historical panel are indexed or relative. Simulation parameters, model counts, reference years and method constants are public methodological quantities rather than internal operating levels.

[HR] Source S1 is a supplied master project registry containing narratives, metrics, decisions, governance records, attribution and source references. The referenced originals were not found in the available local inventory. Its source claims therefore remain registry-attested.

[PD] Reproduction package: the private annex contains the selected extract, exact source locators, source-file hash, full raw calculations and executable Python model. The public calculation file contains only anonymous indexed outputs and explicit scenarios. The source workbook is preserved unchanged.

[AI] Calculations C1–C6 cover variation, versions, selection, economics, information value and power. Analytical rounding sensitivity changes the interpretation of the only process flag. The evaluation’s failure to detect economic break-even under smaller designs is a separate proposed-design insight.

Documentation standard: Luna Sol operating-case skill. The private annex must remain separate from the public case.

REFERENCES / PROCESS & EVIDENCE

Primary sources anchor the analytical methods

References support methods and interpretation; none is evidence of this client's results. Each entry is linked.

1. Hernán, M. A., & Robins, J. M. Causal Inference: What If. Author-hosted text. Potential outcomes, causal diagrams and identification assumptions.
2. NIST/SEMATECH. e-Handbook of Statistical Methods, Individuals Control Charts. Moving-range estimation and individuals limits.
3. Wheeler, D. J. (2012). Exact Answers to the Wrong Questions. SPC Press. Process behavior and the limits of statistical interpretation.
4. Laney, D. B. (2002). Improved Control Charts for Attributes. Quality Engineering. Extra-binomial dispersion and p' chart construction.
5. Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology, 34, 215–220.
6. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. Independent classification and chance-corrected agreement.
7. Collier, D. (2011). Understanding Process Tracing. PS: Political Science & Politics. Hoop, smoking-gun and doubly decisive tests.
8. Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika. Conceptual boundary for comparability; not a requirement to fit factor models to ordinary rates.
9. OECD / European Commission JRC (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. Normalization, weighting and sensitivity.

Public methodological literature. Source identities and raw case records are confined to the private annex.

REFERENCES / EVALUATION & ORGANIZATIONAL LEARNING

Evaluation and learning literature define the unresolved tests

The proposed framework connects measurement discipline to operational learning and allocation decisions.

10. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*. Author manuscript: cohort-time effects and appropriate comparisons.
11. Roth, J. (2022). Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*. Limits of pre-trend tests.
12. Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*. Schedule-specific planning.
13. Hemming, K., et al. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*. Practical rollout and design considerations.
14. Strong, M., Oakley, J. E., Brennan, A., & Breeze, P. (2015). Estimating the Expected Value of Sample Information Using the Probabilistic Sensitivity Analysis Sample. *Medical Decision Making*.
15. The W. Edwards Deming Institute. The Funnel Experiment. Institutional explanation of tampering and common-cause variation.
16. Campbell, D. T. (1976). Assessing the Impact of Planned Social Change. Original occasional-paper record, ERIC. Indicators and distortion under social pressure.
17. Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance implications of strategic performance measurement in financial services firms. *Accounting, Organizations and Society*.
18. Tucker, A. L., & Edmondson, A. C. (2003). Why Hospitals Don't Learn from Failures: Organizational and Psychological Dynamics That Inhibit System Change. *California Management Review*; author working paper.
19. Goodhart, C. A. E. (1984). Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice*, pp. 91–121. Conceptual analogue for relationships that change under control; publisher chapter record and abstract.

Method references are primary papers, author material or original institutional documentation; descriptive case conclusions remain bounded by S1.

The revised case supports a decision, not a claim of proven impact

[AI] Review standard: executive usefulness, methodological validity and reproducibility. This is internal editorial review, not external journal peer review.

Major issue	Revision and final disposition
Privacy inversion	Removed all numerical pass-rate indices. Only complementary failure indices are public; raw values and identities remain private.
Statistical certainty	Added rounding reconstructions. The only MR flag is explicitly provisional and not robust to source precision.
Release criteria	Separated diagnostic validation from the higher evidence threshold for causal credit and financial booking.
Causal coherence	Aligned the reduced executive DAG with expanded mediators, fidelity and measurement paths; removed completer-conditioned effect testing.
Economics and power	Kept scenarios separate from historical outcomes; compared a design-specific MDE with the same-scale break-even; supplied a conditional redesign.

[AI] Minor issues resolved: framework naming, source-version direction checks, contribution-grade wording, a distinct evidence legend and explicit indexing boundaries. The reusable skill now contains a direct check against complementary-index inversion and tests for weak source evidence.

[AI] Referee disposition: suitable as a descriptive operating-mechanism and evaluation-design case. Historical causal impact, realized financial benefit, primary-artifact authentication and field-test feasibility remain unestablished. Those are limits on the case’s claim scope, not effects inferred from missing data.

[AI] Partner critique: a COO can act on page one—commission source-precision and population validation, retain governance cadence, and hold causal and financial credit. The most consequential finding is that adverse score arrows outnumber defensible process signals, and the sole signal itself depends on rounding.

[PD] Next evidence gate: obtain the unrounded event panel, fixed cohort roster, intervention delivery log and mature customer endpoint. Reopen the impact question only after that evidence supports a specified estimand and a test capable of resolving economic break-even.

Editorial review used three delegated workstreams and author revision. Computation, claim-scope checks, confidentiality screening and page rendering support this disposition.