

OPERATING CASE STUDY / OPERATIONAL EXCELLENCE

Enterprise Operating Scorecards

Turn a detected error into a completed correction

An anonymized project case study of a weekly operating tracker: linking performance measures, targets, source versions and action ownership so executives can use the report to make defensible decisions.

PROJECT CONTEXT / DISTRIBUTED SERVICE OPERATIONS

BLUF / BOTTOM LINE UP FRONT

Close the correction loop. Then release the weekly decision.

RECALCULATED FINDING

Correcting a target reduces one reported gap by 45.45% with no change in actual performance. A separate weekly gain fails its endpoint arithmetic. Both findings change how the report should be used.

CONTRIBUTION IN THE SOURCE RECORD

Tracker design, editorial standards and correction direction. Registry-attested ownership; hoop-level support.

EVIDENCE STATUS

Evidence-reconciliation and operating-design case. Adoption, customer impact and realized savings remain unestablished.

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

EXECUTIVE ABSTRACT

Make the correction enforceable at publication, then test the value of the decision

An evidence-reconciliation case about a weekly operating tracker. Its design made diagnosis and action visible; unresolved targets and mixed source versions still changed the decision narrative.

CONTEXT | A recurring read needed an accountable operating explanation

[HR] The register describes a tracker covering process quality, dispatch exceptions and resolution effectiveness across two operating groups. It combined a rolling history, goal, diagnosis and action commitment. The enterprise report's corresponding diagnosis cells were recorded as blank. Later versions added source disclosures and overdue validation items.

OBJECTIVE | Make each released decision reproducible

[PD] Bind the metric, eligibility, reporting vintage, target version and action owner into one approved record. Preserve what was originally published while exposing what a later source changed. Do not let a descriptive movement, a classification hypothesis or a composite weight become unsupported causal direction.

APPROACH | Reconcile the record before designing the control

[RC] Recompute target-gap changes, weekly comparisons, source alternatives and tentative process limits. [AI] Separate detected defects from closed corrections, disclosure from control, and repeated tab agreement from independent evidence. [PD] Evaluate a controlled release protocol using independently adjudicated decision loss per fixed eligible packet.

FINDINGS | The same operations can produce a different management story

[RC] A target correction reduces an earlier reported gap by 45.45% with an unchanged actual. The stored-snapshot resolution comparison yields a 0.17% relative decline; a later-restated prior yields 0.36%. In the other group, the gain calculated from stored endpoints is 23.93% smaller than the source's stated gain. None is evidence of an intervention effect.

DECISION VALUE | Assign a release owner and a zero-unresolved-conflict gate

[PD] The report owner should hold affected decision claims when a material target, definition or arithmetic conflict remains unresolved; unaffected, labeled descriptive data can still circulate. Closure requires a corrected artifact, an updated build specification and traced downstream versions. Scale the protocol only after its decision-loss reduction clears its cost and delay burden.

IMPLICATIONS | Credit the artifact work within its evidence boundary

[HR] The register assigns the principal tracker ownership, editorial direction and correction work. [AI] This is hoop-level support for artifact stewardship. The current case did not reopen the original workbooks; field execution, audience use and realized operating or financial benefit remain unestablished.

S1: selected project register; metric records M01–M14 and linked evidence. Labels: HR historical record; RC recalculated; AI analytical inference; PD proposed design. Identities and original values remain private.

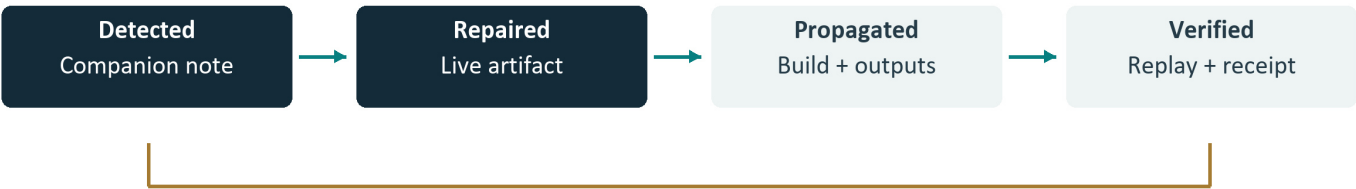
EXHIBIT 1 / CORRECTION CLOSURE

A detected discrepancy remains an open defect until every affected release is controlled

[HR] The register records a target discrepancy in a companion narrative, later correction of workbook cells, and a build specification that still carried the old target.

DETECTION IS THE START OF A CORRECTION

Proposed release control • historical repair stopped short of full closure



Close the issue across artifacts, configuration and dependent releases.

A corrected workbook can coexist with instructions that recreate the old error.

[AI] The failure crossed artifact boundaries. Detection created a note; the note did not create an enforceable dependency on the goal cell or the next build. A manual fix later corrected the visible workbook while the reusable instructions could reintroduce the defect. Detection, repair and prevention are separate states.

Closure evidence	What it must establish
Decision record	Named source authority, approved target version, affected metrics and owner. A disagreement has a disposition rather than an accumulating footnote.
Published artifact	Corrected live cells, aligned narrative and regenerated outputs. Preserve the original as a superseded release with an explicit change record.
Build and recipients	Versioned configuration agrees with the artifact; a replay passes. Dependent reports are identified and corrected or notified through their owners.

MANAGEMENT IMPLICATION

[PD] Count a correction as closed only after the next controlled build reproduces it and downstream disposition is recorded. Detecting the issue or editing one cell does not meet that threshold.

[HR] S1:M07; decisions D02, D04, D07; evidence E02, E08, E13–E14. [PD] Release controls are proposed. Existing delivery recipients and downstream correction completion are unestablished.

EXHIBIT 2 / TARGET INTEGRITY

Correcting the target cuts the reported gap without improving the operation

[RC] Rebase the earlier overstated gap to 100. Hold the actual observation fixed and change only the target to the source-designated authoritative version.

THE GAP MOVES WHILE THE ACTUAL STAYS FIXED

Recalculated • earlier reported gap = 100 • no performance change



45.45% smaller reported gap • 0 change in the actual

Target correction is a restatement of context, not an operating gain.

[RC] The corrected gap is 54.55 on the old-gap index, a 45.45% reduction. The source’s statement that the old gap was 83.33% larger uses the corrected gap as its denominator; the two calculations answer different questions. Neither calculation is an operating performance gain.

[HR] The register describes a variance back-solve and a later goal-table tie-out. These are distinct consistency checks, available here as transcriptions. Original formulas and source-ownership approvals were not recovered.

Tempting conclusion		Defensible interpretation	
The most weighted metric causes the largest ranking error.		Weight is only one input. The scoring transform, direction, scale, caps and regional distribution determine how a target error changes ranking.	
Correct the target and credit the gap reduction.		Correct the target and restate the decision context. Keep operational actuals unchanged and label the movement as a measurement correction.	
Every old release is now corrected.		A current workbook correction does not establish that prior decks, narratives or build instructions were updated.	

RELEASE RULE

[PD] Store a governed target identifier with its effective period. Fail the affected decision release when the target configuration and the declared source disagree.

[RC] S1:M07, private/raw_calculations.py. Public gap indices preserve confidentiality; raw targets and absolute errors are private. Source-designated authority is not independent authentication.

EXHIBIT 3 / REPORTING VINTAGES

A weekly comparison changes when its prior period comes from a different vintage

[RC] The same current resolution value produces different changes against the stored prior and the later-restated prior.

TWO LEGITIMATE QUESTIONS NEED TWO LABELED BASES

Relative Group A resolution change • same current value • different priors

Published-snapshot comparison
−0.17%

Later-vintage comparison
−0.36%

$$\text{Current} - \text{stored prior} = (\text{current} - \text{revised prior}) + (\text{revised prior} - \text{stored prior})$$

Identity is in level units; percent comparisons have different denominators.

Keep the publication history. Add a same-vintage comparison and a revision bridge.

[AI] A published-snapshot comparison measures change from the value readers saw last time. A same-vintage comparison uses the current source for both periods. An append-only history protects the audit trail; it does not make those questions identical.

[RC] For the registered Group A resolution values, the published-snapshot comparison is −0.17% and the later-vintage comparison is −0.36%, both relative changes. The original narrative uses the latter magnitude while describing the former reference. The movement remains negative under both; its size is version-sensitive.

Required field	Reason to retain it
Observation period and extraction vintage	Separates a new operating period from a revision to a prior period.
Published value and current comparable value	Preserves the historical communication while enabling a same-definition comparison.
Revision reason and source lineage	Distinguishes verified maturation, recoding, coverage changes and unresolved source disagreement.

ANALYTICAL CONTRACT

[PD] Display both comparison bases when material. Freeze a single source vintage for process charts. Do not attribute every small source difference to maturation without a revision log.

[RC] S1:M05 and E12–E14. Percentages rounded for publication; exact inputs and alternative endpoint calculations remain in the private annex. These are descriptive comparisons, not process signals.

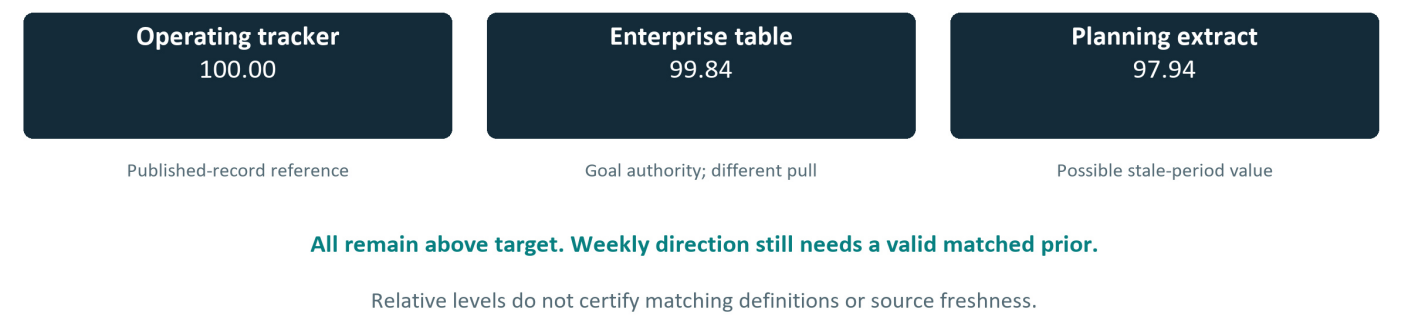
EXHIBIT 4 / SOURCE ALTERNATIVES

Keep the source fork visible when it changes the interpretation of a weekly move

[AI] Select a source for a stated use, then show the decision sensitivity. Do not select whichever value makes the operating story cleaner.

SOURCE ROLE PRECEDES SOURCE SELECTION

Same labeled closing period • tracker level = 100 • vintage comparability unresolved



[HR] Three registered dispatch-exception values were assigned to the same closing period in different artifacts. The planning extract’s value may reflect a stale period; its label is not proof of comparable timing. The operating tracker is the record of what it published, while the enterprise table is the source designated for goals. Those roles do not make either universally authoritative.

Reconciliation finding	Decision consequence
Dispatch exception levels differ across artifacts.	All registered alternatives remain above the standing target, but a weekly direction requires a valid prior from the same population and vintage. Do not splice the planning baseline into a current operating series.
Group B’s reported weekly resolution gain does not reproduce.	The stored endpoints imply a gain index of 76.07 when the narrative claim is indexed to 100. The restated and enterprise endpoint pairs also fail to reproduce the claim.
Two tabs or reports agree.	Agreement is a useful tie-out; shared upstream data can create correlated errors. Verify lineage before calling the agreement independent corroboration.

PUBLICATION DISPOSITION

[PD] Correct the reproducible arithmetic; retain unresolved source forks. The analysis supports direction only where the comparison contract survives the alternative.

[HR/RC] S1:M03, M05–M06; E13–E15. Tracker reference = 100 for this same-period comparison. The diagram does not imply all sources share a certified vintage or construct.

EXHIBIT 5 / PROCESS EVIDENCE

The available weekly histories do not establish a process shift

[RC] Tentative individuals and moving-range limits are calculated for the six same-source histories transcribed in the enterprise table.

Assessment	Result and implication
Ordered observations	Each short history contains six weekly observations on its stated basis. Earlier proxy observations are excluded from the later process-quality series.
Tentative limits	No individual point crosses its limit. Group A sorting’s last moving range crosses its MR limit; plausible rounding can reverse this classification.
Interpretation	The short baseline, rounding and a longer mixed-vintage alternative do not support a robust signal label. Investigate the movement; do not label it established routine variation or improvement.
Rates and counts	No complete matched event-count and exposure panel exists for these histories. A Laney p’ chart cannot be fitted from rounded rates or a single period’s counts.
Next decision	Collect a longer same-definition baseline with mature counts, exposure and version logs; freeze phase boundaries and rules before interpreting subsequent signals.

[AI] A change between the last two observations is descriptive arithmetic. A target miss answers whether performance meets a standard; a process signal answers whether the sequence is unusual relative to a specified process. Neither is a causal treatment effect. A large target miss can coexist with routine variation.

[PD] Preserve missing periods as missing. The register distinguishes an absent weekly tracker from a prior-period value later recoverable in a source pull. The later value cannot recreate an original publication snapshot. Do not insert a repeated value to make the chart continuous.

EXECUTIVE USE

[PD] Use the short histories to identify what requires investigation and better measurement. Do not narrate within-limit weekly direction as established improvement or deterioration.

[RC] S1:E14; private/raw_calculations.py applies XmR mean ± 2.66×mean(MR). Limits are tentative and based on rounded transcribed values. No historical signal claim or Laney p’ result is asserted.

EXHIBIT 6 / ACTIONABLE MEASUREMENT

An adoption-rate collapse can coexist with a larger recorded numerator

[RC/AI] The recorded denominator change creates a testable accounting question before any claim about frontline behavior.

THE DENOMINATOR CHANGES THE BEHAVIORAL STORY

Conditional accounting • each prior quantity separately indexed to 100

Rate index
32.55

×

Exposure index
354.10

÷ 100 =

Implied numerator
115.26

A lower rate does not establish a lower action count.
The identity requires compatible rate, exposure and count definitions.

Changed eligibility can invalidate behavioral comparisons even when the arithmetic holds.

[RC] With each prior quantity indexed to 100, adoption falls to 32.55 and eligible exposure rises to 354.10. Their product implies a numerator index of 115.26 if definitions and counts are linked. Original counts are unavailable to verify that condition.

[AI] The rate alone does not establish that fewer actions occurred. Nor does the conditional numerator show improved adoption: broader eligibility can change the construct and population. Counts under different definitions need not measure the same behavior.

Observed or asserted pattern	Evidence required before action
One category rises as another falls.	Event-level transitions, unchanged eligible exposure and coding-version logs before calling it a tagging artifact. A paired swing supports a validation hold, not proof of cause.
A low-performing cohort rebounds.	Frozen membership, comparable untreated histories and selection-aware analysis before crediting the action.
Resolution rises after a narrower action list.	Delivery records and a controlled comparison. Different operating groups are not a natural experiment merely because they chose different actions.

[PD] Hold attribution until definitions reconcile. Assign a validation owner and deadline, with the decision consequence stated.

[RC/AI] S1:M03/M09 denominator-change statement; D08–D09. Three separate prior indices each equal 100; multiplication uses the rate identity, not pooling. The implied numerator is conditional, not a recovered historical count.

EXHIBIT 7 / CONTROLLED RELEASE

Release the decision only when its measurement and action commitments agree

[PD] A governed tracker is a chain of explicit decisions with one accountable release owner.

Gate and owner	Required evidence	Release consequence
Source / data owner	Metric label, unit, eligibility, missingness and source vintage match the declared contract. Proxy scope is explicit.	Hold the affected inference if definitions cannot be reconciled.
Comparison / analyst	Target authority, period, endpoint arithmetic, source fork and appropriate variation screen agree.	Release the declared descriptive comparison; suppress unsupported trend or causal wording.
Action / operating owner	One testable mechanism, owner, due date, measurable target and verification route. Proposal versus approval is labeled.	Commit only the authorized action; preserve overdue status across reporting cycles.
Publication / report owner	Artifact, configuration, correction log and dependent outputs use the approved version.	Publish with a version ID; a material defect reopens the release.
Value / sponsor and finance	Independent decision-loss evaluation, delay/coverage components and incremental cost.	Scale only inside the validated economic and operating boundary.

[PD] Test the release protocol through randomized rollout across publication teams with a frozen inventory of eligible decision packets. Charge both erroneous releases and unresolved packets at the deadline under a prespecified loss schedule. Report unweighted error, hold, coverage and delay components so the composite cannot conceal a costly tradeoff.

COO ACTION

[PD] Commission the versioned release protocol and an independent evaluation. The first approval is for a controlled pilot with explicit economics; customer-outcome and savings claims require separate evidence.

[PD] Proposed decision charter informed by S1’s recorded defects. Public case does not claim the protocol was deployed or that the tracker itself changed field outcomes.

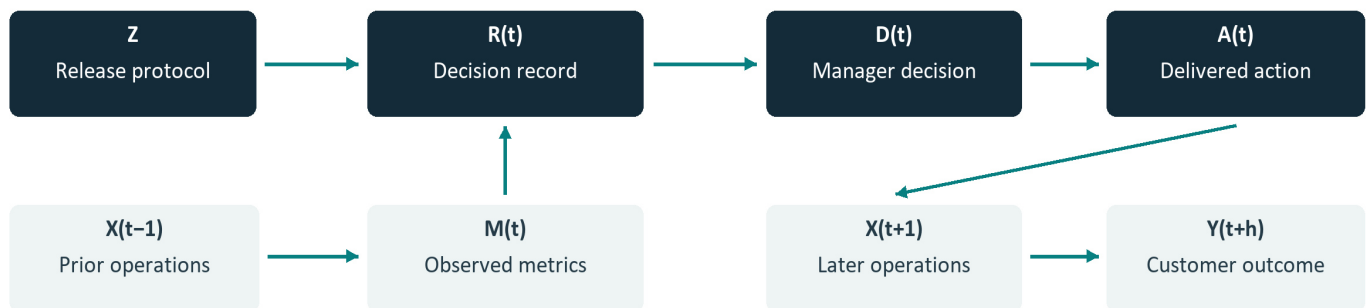
TECHNICAL FOUNDATION / THEORY OF CHANGE

Better reporting can change decisions without yet establishing a customer effect

[PD] The causal graph separates publication governance, managerial use, delivered action and later outcomes. Time indices keep the operating feedback loop acyclic.

SEPARATE THE REPORTING INTERVENTION FROM FIELD IMPACT

Proposed, time-indexed DAG • arrows are hypotheses, not estimated effects



Q: measurement regime $\rightarrow$ $M(t)$ • G: target version and V: source vintage $\rightarrow$ $R(t)$

C: demand, staffing, node mix and concurrent changes $\rightarrow$ decisions, actions and outcomes

Independent adjudication of $R(t)$ identifies decision loss $K(t)$, the proposed primary endpoint.

Later operations feed a future reporting cycle; they do not close a same-period causal loop.

[PD] The release protocol Z changes the decision record $R(t)$. Managers may use $R(t)$ to change $D(t)$, which can alter delivered action $A(t)$, subsequent operating state $X(t+1)$ and mature customer outcome $Y(t+h)$. Prior operating state $X(t-1)$ generates the measured inputs $M(t)$ used in the current report. Observing a changed report does not identify the downstream path.

[PD] Measurement regime Q changes the observed metric. Target version G and source vintage V change what the record reports. Named common causes C include demand, staffing, node mix and concurrent operating changes. They can affect managerial choices and field outcomes. Additional unmeasured paths remain possible; residuals cannot identify their origin.

[PD] The proposed rollout randomizes protocol availability across publication-team clusters. Independent blinded adjudication assesses decision-loss burden $K(t)$ on an upstream-frozen set of eligible packets, including withheld packets at the deadline. It tests the reporting intervention's total assignment effect on decision loss, without conditioning on post-treatment release.

[AI] The source supports descriptions of record design and correction. It does not establish manager use, verified action delivery or the causal path to customer outcomes. A downstream customer-effect test would require a separate maturity horizon, exposure panel, spillover design and sample-size justification.

[PD] Proposed DAG and reporting-policy evaluation. G , V , Q and C arrows are summarized in labeled prose for readability. S1 supplies failure modes, not identified path coefficients.

TECHNICAL FOUNDATION • DECISION PATHWAY AND ESTIMANDS

A better decision record can reduce error before it changes operations

[PD] Evaluate the tracker at the decision boundary it controls; treat customer impact as a separate downstream claim.

[PD] Use time-indexed nodes: $Z \rightarrow R(t) \rightarrow D(t) \rightarrow A(t) \rightarrow X(t+1) \rightarrow Y(t+h)$. Prior operations $X(t-1)$ generate $M(t)$; Q affects measurement, while $M(t)$, authorized target G and source vintage V inform $R(t)$. C includes demand, staffing, node mix and concurrent changes; U denotes unmeasured causes. Later operations enter a future reporting cycle, not a same-period directed cycle.

Quantity	Population, contrast and boundary
Tracker-policy effect	[PD] ITT on independently adjudicated decision-loss burden per frozen eligible metric-week packet, governed release versus current workflow, at a preregistered deadline. Retain withheld packets in the denominator.
Required components	[PD] Report invalid releases, false holds, readiness, coverage and delay separately. Freeze exhaustive, mutually exclusive statuses and loss weights before assignment; a composite cannot hide a failing component.
Operating-action effect	[PD] Effect of a specified delivered package on eligible operating opportunities at a declared horizon. Assignment to a better tracker is not identical to receiving or executing the package.
Customer effect	[PD] Difference in mature, consistently defined customer outcomes under tracker assignment. Requires evidence across the downstream pathway; it is unpowered and unidentified here.

[AI] A changed narrative, declared owner or completed workbook supports a record-level claim. It does not establish readership, changed decisions, execution or customer benefit. Conditioning on adoption or completed actions would change the policy estimand and may introduce selection. Residuals cannot identify which unmeasured pathway generated them.

Source S1: registered tracker design and release history. Causal edges and prospective estimands are proposed; historical downstream impact remains unestablished.

The metric, denominator, target and source vintage must travel together

[PD] A Comparable Decision Record stores the value and the conditions under which a comparison is valid.

Record element	Validity and invariance requirement
Sorting compliance	[PD] Fix eligible processing events, pass definition, node mapping and aggregation weights. Do not join an earlier proxy to a later operational measure as one unchanged series.
Dispatch exception	[PD] Bind exception numerator to dispatched exposure, category rules and deduplication. A falling rate need not mean fewer exception events when exposure changes.
Resolution effectiveness	[PD] Match resolved cases to eligible anomalies and their cohort. Separate overall from stage-specific rates. Record maturation, carryovers and revised prior-period values.
Target and population	[PD] Version authorized goals and effective dates. Node type is not automatically business-line identity; one matching aggregate does not prove set equivalence or future invariance.
Action commitment	[PD] Record owner, due date, measurable target, evidence status and closure rule. Written intent, accepted assignment and verified completion are different constructs.

[PD] Preserve both an as-published snapshot and a comparable latest-vintage series. Cross-vintage change equals same-vintage change plus revision of the prior period, provided definitions and population remain comparable. A missing source file is a retrieval issue until alternate-source reconciliation establishes whether the metric itself changed.

[PD] Begin with 600 independently double-coded packets stratified by node, metric, revision and disputed status; apply a frozen codebook blind to treatment. This gives about ± 5 percentage points for a worst-case agreement proportion at an assumed design effect of 1.5; kappa precision additionally depends on prevalence. Pilot the kappa interval and expand if needed. Adjudicate after independent coding; report confusion matrices, prevalence, agreement and kappa with team-clustered uncertainty.

[AI] Clear prose must retain the distinction between a measured driver, a plausible mechanism and an unresolved explanation.

Source S1: row-map changes, proxy definitions, target conflict, snapshot revisions and action fields. Validity and classification protocols are proposed.

Weights alone cannot rank misses measured in different units

[AI] The register supplies component weights but does not recover the complete historical normalization, aggregation and ranking contract.

[PD] A composite requires orientation, transformation, reference distribution, target treatment, weights, missing-data rules and aggregation. Define whether a high score denotes harm, distance from target or investment priority. A target miss is neither an avoided-loss estimate nor a causal treatment effect.

Synthetic counterexample	Candidate A	Candidate B
Two raw gaps	8 and 2	2 and 8
Weights 0.60 / 0.40	Score 5.60	Score 4.40
First unit multiplied by 0.01	Score 0.848	Score 3.212

[RC] The larger-gap priority changes from A to B although the physical states are unchanged. This invented example proves that an unnormalized weighted sum is unit-dependent; it does not reconstruct or condemn the historical ranking algorithm.

[PD] Freeze a dimensionless transformation and test unit invariance, goal-revision sensitivity, missing-component behavior and weight uncertainty. Track rank intervals, ties and transitions under plausible rules. Do not create a point ranking when reasonable specifications reverse the priority.

[PD] Preserve a noncompensatory guardrail where one safety or customer condition must not be offset by a favorable unrelated component. Finance should separately price a deliverable intervention, feasible response and marginal loss avoided; largest weighted shortfall need not offer the highest return.

[AI] Indexing protects raw scale but does not make independently normalized metrics economically interchangeable. A revised goal can shrink the reported gap without any change in operating performance.

Source S1: weight-aware narrative and goal mismatch. Synthetic unit-invariance example is reproduced in technical_sensitivity.py. OECD/JRC composite-indicator methods inform the proposed tests.

TECHNICAL FOUNDATION • PROCESS VARIATION

A weekly direction is not automatically a process signal

[PD] Classify only commensurate observations within a declared phase and source vintage.

[HR] The register supplies six consecutive weekly observations for each input and network from one later enterprise extract. Earlier proxy periods and mixed-vintage tracker values are separate. Eligible weekly denominators and within-week distributions are unavailable.

[PD] For a commensurate series, MR is the absolute adjacent difference; tentative individual limits are $\text{mean}(x) \pm 2.66 \times \text{mean}(\text{MR})$, and MR upper limit is $3.267 \times \text{mean}(\text{MR})$. The rounded factor 2.66 approximates $3/1.128$. A linear change of units preserves classification. Sorting is shown only as non-pass burden, consistent with the earlier public case; publishing the pass-rate complement would expose confidential levels.

[RC] No individual point in the six enterprise histories crosses its tentative individual limit. Group A sorting's terminal moving range crosses the MR limit. An extended, stitched history also flags the terminal individual point, but its vintage and phase assumptions differ. Rounding scenarios can change both classifications. The conclusion is investigate and reconcile, not certified signal or certified stability.

[PD] Investigate phase changes, denominator volatility, revisions, autocorrelation and case mix before treating limits as stable. With only six observations, one moving range can dominate the estimate. Recalculate under credible rounding and source-version alternatives; hold the classification when the result changes.

[PD] Future counts permit rational-subgroup p charts. For large denominators with extra-binomial variation, evaluate Laney p' using standardized subgroup residuals and their adjacent moving ranges. Counts, denominators and enough subgroups are prerequisites; large volume alone does not justify the adjustment. Never infer exact event counts from rounded rates.

[AI] Within-limit movement alone does not establish improvement or deterioration. Two-point comparisons elsewhere have non-diagnostic limits in the private annex. A goal breach concerns a management threshold, distinct from a process limit.

Source S1: single-vintage six-point series. Tentative process analysis is descriptive and conditional; no treatment effect is estimated.

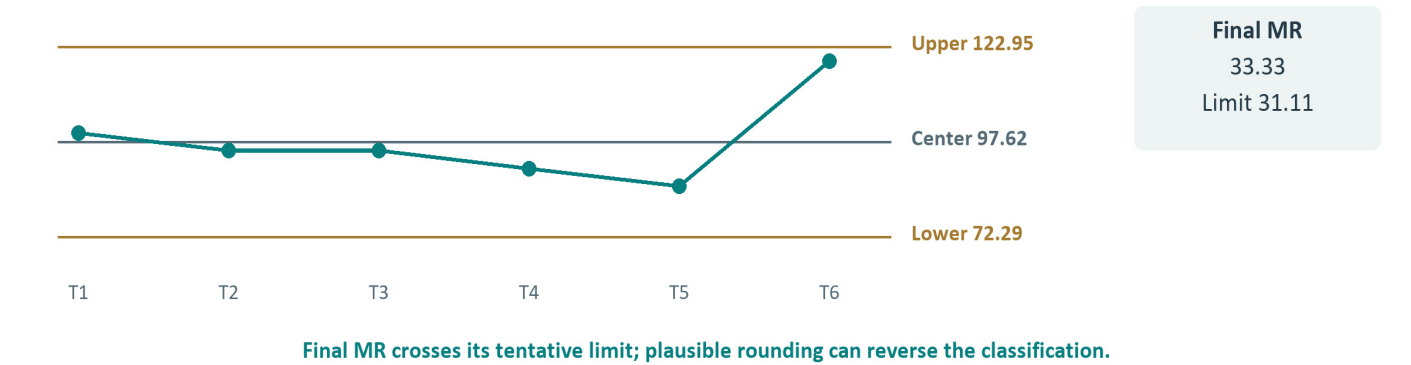
TECHNICAL EXHIBIT / VARIATION SENSITIVITY

A tentative signal depends on the window and the precision of the source

[RC] A short same-vintage series and its moving ranges produce different screens; extending the history changes the answer again.

POINT AND MOVING-RANGE SCREENS ANSWER DIFFERENT QUESTIONS

Group A sorting non-pass burden • first period = 100 • tentative six-point limits



Registered-data analysis	Permitted interpretation
Same-vintage six-point history	Individual screen: no crossing. Terminal moving-range screen: a crossing in one sorting series; the other five histories have no crossing.
Extended reporting history	The corresponding terminal individual point crosses a limit, but earlier observations mix release vintages; another group’s early proxy values also break comparability.
Rounding sensitivity	Witness perturbations inside the stated display precision can reverse the terminal classification. These witnesses establish non-robustness; corner searches are not proofs of global extrema.

[AI] The source narrative calls the final directional movement real after a source tie-out. The tie-out can support continuity of the reported series. It does not certify a process shift. More baseline periods, unchanged definitions, underlying precision and event/exposure records are required before escalating on a statistical rule.

[RC] S1:E14 and registered earlier snapshots; private/raw_calculations.py. First non-pass burden = 100. Limits and MR are in that same index; raw pass levels and alternative histories are private.

TECHNICAL FOUNDATION • SELECTION AND MODERATION

Bottom-ranked cohorts can appear to recover without an intervention

[AI] Selecting weak units changes the expected follow-up even when the underlying process is unchanged.

[PD] Freeze the cohort and selection rule before follow-up. Distinguish improvement among originally selected units from a refreshed ranking that selects different units each week. Preserve non-focus comparisons, eligibility changes, missing records and the original selection score.

Synthetic reliability r	Expected rebound from shortfall $d = 12$
0.25	9.0 normalized points
0.50	6.0 normalized points
0.80	2.4 normalized points

[PD] Assume stationary classical Gaussian repeat measures: $X = T + \text{error}$, with independent equal-variance errors. Reliability r is stable between-unit variance divided by total variance. For a selected score d below the population mean, expected rebound is $(1-r)d$. These values are invented sensitivity inputs, not an estimated correction to the tracker’s cohort.

[PD] Estimate reliability from repeated untreated observations, then compare fixed focus and non-focus cohorts on baseline levels, trends and comparable follow-up. If ranking rules or source vintages change, recreate selection consistently before estimating reversion.

[PD] Treat node type, network, volume, stage, staffing and instrumentation as moderators. Define mechanisms and interaction tests; inspect common support. A narrower action list in one network and a favorable week do not identify the effect of focus. Different subgroup significance levels do not establish different effects.

[AI] A staff-to-plan association cannot by itself separate capacity, case difficulty, local management or selective deployment. The tracker can prioritize a controlled test without declaring headcount unnecessary.

Source S1: bottom-cohort reporting and cross-network action narratives. Selection sensitivity is synthetic and executable; historical regression-to-the-mean magnitude is unidentified.

TECHNICAL FOUNDATION • PERSONAL CONTRIBUTION

Attribute the tracker’s decision architecture without claiming its downstream impact

[HR] The register attributes editorial direction, format decisions and ownership roles to the principal; implementation and consequence require separate evidence.

Contribution claim	Process-tracing grade and boundary
Audience and reporting grain	[HR] A session-associated memory record attributes network-grain and executive-read directives to the principal. [AI] Provisional registry-level hoop support; it does not uniquely authenticate the originating human instruction.
Named owner on companions	[HR] The register describes an owner byline on explanatory documents. [AI] Provisional hoop support for the stated responsibility, stronger than mere meeting presence; a byline does not prove sole text or code authorship.
Disclosure and repeatability	[HR] Successive artifacts contain source-gap blocks and a codified workflow. [AI] Their existence supports the design’s recorded evolution; personal authorship and faithful reuse require provenance.
Correction stewardship	[HR] The register describes target reconciliation and withdrawal of a source-break conclusion. [AI] Provisional hoop support for attributed review; locate timestamped decisions and actual corrected releases before stronger grading.
Readership, adoption and benefit	[AI] Distribution, executive use, changed field decisions and customer benefit are unestablished. A completed artifact is not evidence that its recommendations were executed.

[PD] A hoop test checks a necessary condition and is not sufficient proof. A smoking gun directly documents the particular action and distinguishes plausible alternatives. Doubly decisive evidence additionally excludes those alternatives. No stronger grade is assigned from register excerpts alone.

[AI] The source also describes a stale reusable workflow that could reintroduce corrected values. Codification is evidence of repeatability intent, not proof of maintained correctness. The relevant contribution is governing revisions through the final artifact.

Source S1: memory, owner bylines, workflow and correction descriptions. Attribution is registry-based; operational originals and human authorship are not independently authenticated here.

TECHNICAL FOUNDATION • PROSPECTIVE EVALUATION

Test whether governed releases reduce decision loss without creating silent holds

[PD] Randomize the release workflow at a level that contains shared authors, reviewers and approval behavior.

[PD] Define one eligible metric-week decision packet before assignment. Freeze deadline, material-error rules and an independent reference ledger. A blinded review team classifies every packet, including withheld and unresolved records. Prevent the release team from choosing the evaluation denominator or adjudicating its own outcome.

Design element	Required contract
Assignment	[PD] Randomize publication teams to stepped rollout sequences. Account for common calendar shocks, team correlation, learning, anticipation and interference; preserve assignment in the policy ITT.
Decision-loss outcome	[PD] Preregister exhaustive statuses and loss weights. Distinguish material erroneous release, justified hold, false hold and unresolved-at-deadline states. Do not count the same delay twice.
Guardrails	[PD] Report unweighted error, false-hold, readiness, coverage, cycle time and reviewer workload alongside the primary burden. An average gain cannot conceal unacceptable suppression.
Investment rule	[PD] Match effect, eligible volume, loss valuation and incremental review cost. Estimate covariance and policy compliance from a pilot; compare design-specific precision with economic break-even.

[PD] Calibrate finite-cluster inference and power for the actual sequences, periods, covariance and economic threshold. Freeze the primary endpoint and multiplicity plan. Use the synthetic model as a planning illustration, not as evidence of achieved precision or approved staffing.

[PD] If adoption is nonrandom, a group-time comparison requires defensible untreated timing, overlap and pre-trend diagnostics; the randomized model does not power that alternative. Customer outcomes need a separate mature-outcome design and are not valued here.

[AI] The historical record does not establish that the tracker changed decisions or losses. The prospective test is the route to that claim.

Proposed publication-team evaluation. Economic categories and statistical assumptions are synthetic until calibrated; customer effects remain separate.

TECHNICAL FOUNDATION • INTELLECTUAL CONTRIBUTION

The Comparable Decision Record turns a polished narrative into an auditable decision

[PD] CDR binds metric, population, source vintage, target, evidence class and action commitment in one versioned record.

[AI] OECD/JRC composite-design guidance motivates explicit normalization and sensitivity. Ittner–Larcker–Randall motivate testing performance-measurement systems against results. Campbell’s indicator critique warns against substituting a scored proxy for purpose. Tucker–Edmondson’s organizational-learning lens distinguishes recording a problem from changing the system that reproduces it.

Falsifiable proposition	Test and disconfirming observation
P1 • Revision visibility	[PD] Splitting same-vintage movement from historical revision should reduce materially incorrect explanations. An interval excluding the prespecified worthwhile reduction contradicts the practical claim for the tested workflow; nonsignificance alone does not.
P2 • Unit invariance	[PD] A defensible priority rule should preserve ordering under unit conversion alone. A changed ordering with unchanged physical states rejects that rule’s claimed comparability.
P3 • Correction closure	[PD] Binding a validated correction to templates and release tests should reduce recurrence versus disclosure alone. A confidence interval excluding the prespecified useful reduction, under comparable exposure, contradicts the practical prevention claim.
P4 • Honest specificity	[PD] Explicitly labeling hypotheses while retaining owner, deadline and measurable target should preserve readiness without increasing unsupported causal statements. A material readiness loss or higher unsupported-claim incidence contradicts that joint prediction.

[AI] The case extends these ideas through a versioned decision contract: a clear sentence can remain wrong when its comparator, target or source lineage is incompatible. A target revision can improve the reported gap without improving operations; an unchanged template can reproduce an already disclosed mismatch.

[AI] These are analytical implications and proposed tests. The case does not demonstrate improved learning, reduced gaming or realized decision-loss reduction.

Primary composite, performance-measurement and organizational-learning literature supports the interpretation. CDR and its propositions are proposed, not externally validated.

TECHNICAL FOUNDATION · LIMITATIONS AND REPRODUCIBILITY

The tracker is reproducible only when unresolved conflicts remain visible

[AI] This register-based case supports analytical reconciliation and a governed-release design; downstream impact remains unestablished.

Boundary	Consequence and required record
Source recovery	[AI] Register excerpts describe originals; this case does not authenticate the full workflow or distribution. Keep source hashes, locators, versions and extraction scope privately.
Comparability	[AI] Source vintages, proxy boundaries, missing snapshots, row maps and goal effective dates remain incomplete. Do not average incompatible values or assign the discrepancy to pull timing without evidence.
Statistical inference	[AI] Short aggregate series and absent weekly counts limit stable process classification. Rankings, fixed-cohort outcomes and site histories are unavailable; selection sensitivity is not an estimated correction.
Mechanism and attribution	[AI] Paired category swings and staffing contrasts do not identify tagging error or an operating cause. Contributor roles are registry-attested; field execution and customer results are undocumented.
Economic consequence	[AI] No realized savings or measured historical information value is established. Proposed loss weights, covariance and cost parameters require calibration before investment approval.
Privacy and replication	[PD] Retain raw values, identities and transformation keys privately. Publish one numeric view for complementary rates; test companion artifacts for inversion and identity leakage.

[PD] Replication order: freeze the source manifest; reproduce the metric dictionary and target authority; reconcile same-period versions; verify numerator/denominator identities; calculate single-vintage series and rounding sensitivity; test normalization and ranks; rerun scenario models; inspect every rendered artifact and active reusable template.

[PD] Preserve competing credible readings and their decision consequences. A proposed driver remains a hypothesis until a discriminating test supports it. A correction is closed only when affected release surfaces, downstream copies and reusable instructions are updated and verified.

[AI] Current replication covers register arithmetic and explicit synthetic scenarios. It does not reconstruct unavailable event-level truth or establish real-world uptake.

Source S1 and internal methodological review. Raw evidence, calculations, identity crosswalk and unresolved discrepancies belong in the private reproducibility annex.

MODEL 1 / PREREGISTERED DECISION LOSS

A valid release policy must price the work it withholds

[PD] All figures on these four model pages are synthetic. They specify a testable planning model, not a historical benefit estimate.

Terminal status at deadline	Illustrative loss weight
Valid released packet	0.00
Materially invalid released packet	1.00
Justified hold with documented reason	0.20
Unjustified hold	0.60
Unresolved or unclassifiable packet	0.80

[PD] Freeze eligible packets upstream. Measure status five business days after packet close and finalize blinded adjudication within 30 days. Released records are classified valid or invalid; unreleased records are justified or unjustified holds. Insufficient evidence in either branch is unresolved. Each packet receives exactly one state.

[PD] Primary burden is the sum of state loss weights divided by all eligible packets, normalized to an independently measured control baseline of 100. Positive benefit is control minus protocol burden. Publish unweighted frequencies, coverage, time to decision and reviewer effort; calibrate weights with stakeholders before randomization.

[PD] Pilot assumptions are equal eligible volumes and baseline burden across team-periods. The 30-day review must use an independent reference ledger and blinded coders; changing labels after seeing allocation invalidates the comparison. Unresolved packets remain in the denominator rather than disappearing.

INVESTMENT BOUNDARY

[AI] This endpoint measures decision and rework burden. It neither proves that an operating recommendation was executed nor establishes a customer outcome.

Synthetic model.py. Loss weights, deadlines and status hierarchy are proposed; no client packet counts or observed loss rates are asserted.

MODEL 2 / ECONOMICS AND SENSITIVITY

The trial must resolve a 6.54% reduction to clear the base-case cost

[PD/RC] Use a synthetic gross decision-loss pool of 100 over one defined implementation horizon.

[PD] Net value = $100 \times r \times a \times (1-o) - K$. Here r is the policy ITT reduction in the same deadline loss endpoint; avoidability $a = 0.85$, overlap $o = 0.10$ and incremental implementation cost $K = 5$. All inputs are assumptions. The value pool and endpoint use the same eligible packets and horizon.

Assumed loss reduction	a = 60%	a = 85%	a = 100%
4%	-2.84	-1.94	-1.40
8%	-0.68	+1.12	+2.20
12%	+1.48	+4.18	+5.80

[RC] Break-even = $K/[100a(1-o)] = 6.5359\%$ relative reduction. The grid holds cost and overlap fixed; its cells are net normalized value units, not money or confidence intervals. A 12% reduction yields +4.18 units at the base avoidability assumption.

[PD] Holding and delay penalties are already in the endpoint. Incremental cost covers extra implementation, tooling and review effort only. If loss weights already represent avoidable, uniquely attributed resource loss, set a and $(1-o)$ to 1; otherwise multiplying again understates benefit. Book cash only when the resource-release mechanism is demonstrated.

[PD] Calibrate the joint distribution of effect, cost, weighting and avoidability; the grid does not establish independent uncertainty. Unequal packet volumes or baseline burdens require an explicit packet-weighted target and revised power.

Synthetic economic scenario. Source record does not establish attributable realized financial benefit or actual policy costs.

MODEL 3 / STEPPED ROLLOUT AND ECONOMIC POWER

A larger rollout can resolve the investment question under stated assumptions

[PD] Randomize teams equally to four switch sequences; ten weekly periods, switches in periods 3, 5, 7 and 9.

[PD] Fit cluster and calendar-period effects with a constant immediate protocol effect. For each team, covariance is $\sigma^2[\rho 11' + (1-\rho)AR1(\phi)]$: $\sigma = 16$ index points, persistent variance share $\rho = 0.20$, serial parameter $\phi = 0.30$. Teams are independent; publication staff must not span intervention boundaries. The known covariance and normal approximation are planning assumptions.

Teams	Zero-null MDE	Economic power at 12%
32	8.64	42.5%
64	6.11	70.7%
96	4.99	86.6%

[RC] MDE targets 80% power against zero at two-sided 5%. Economic power is the probability that the lower 95% bound exceeds the 6.54% break-even when the true reduction is an assumed 12%. The minimum balanced base design is 84 teams; 96 provides additional margin.

[RC] If σ rises to 20, 96-team economic power falls to 68.9%; 128 teams are required under that alternative. Serial correlation is tested separately in the package; correlation does not universally inflate this contrast's variance.

[PD] Pilot team-period covariance, unequal volumes, onboarding lag, missingness and contamination before selecting a final design. Calibrate finite-cluster inference and event-time heterogeneity; changing the constant-effect estimand requires new power. Check baseline trajectories and anticipation, but a nonsignificant pre-trend is not proof of comparability.

Synthetic cluster/time GLS with covariance sensitivity. Full-design and profiled information calculations agree. Historical impact and achieved trial power are unestablished.

MODEL 4 / INFORMATION VALUE

A diagnostic audit is worthwhile only when it changes the best next investigation

[PD] This information decision is separate from rollout benefit and does not assume a positive customer effect.

Diagnostic action	Target defect	Version defect	No defect
Target reconstruction	+12	−4	−6
Source reconstruction	−5	+14	−6
Retain current review	0	0	0

[PD] Prior state probabilities are 0.45, 0.35 and 0.20. Payoffs are hypothetical diagnostic resource units, including reconstruction effort and delay. A preliminary audit produces one of three signals: correct with probability q , each wrong signal with $(1-q)/2$. Choose the action with greatest posterior expected payoff.

Signal accuracy q	Gross EVSI	Net after audit cost 2
50%	1.238	-0.763
75%	4.169	+2.169
90%	6.168	+4.168

[RC] EVSI is expected optimal payoff after information minus optimal payoff before information. EVPI is 7.500; every scenario stays below this bound. The package includes posterior probabilities and alternative audit costs. These ranges are sensitivities, not estimated information performance.

[PD] Replace priors, diagnostic accuracy and payoffs with pilot evidence. The source documents corrections and validation holds, but does not measure their historical economic consequences. Never add this illustrative diagnostic value to the operating policy value when both concern the same prevented error.

Synthetic finite-state decision model; conditional probabilities and costs are declared. EVSI computed by posterior optimization and independently by joint payoff enumeration.

The deliverable is a corrected decision contract with a testable value proposition

[AI] The case credits the reporting work supported by the register and distinguishes it from a verified operating transformation.

Claim class	Meaning
[HR] Historical record	A statement or artifact is described in the supplied register. Original operational workbooks were not recovered in the targeted source search.
[RC] Recalculated	Arithmetic from registered values or a declared synthetic model; the relevant input class is specified.
[AI] Analytical inference	A reasoned interpretation or constraint; not an identified treatment effect.
[PD] Proposed design	A future protocol, test, scenario or framework. It is not claimed as implemented.

[AI] The derived findings change how to use the tracker: a target correction changes the gap without changing actuals; incompatible vintages change the weekly narrative; a reported gain does not reproduce; and tentative variation flags are sensitive to window and precision. These findings favor a controlled release and evaluation, rather than an unsupported claim of operating benefit.

[AI] Wang–Strong broadens data quality beyond numerical accuracy to fitness for use. W3C provenance represents versions, derivation and responsibility without certifying truth. Shewhart–Deming cautions against reactions to noise. Campbell–Goodhart cautions that scored indicators can change behavior; no gaming is demonstrated here. Tucker–Edmondson distinguishes a local workaround from preventing recurrence.

[AI] This case illustrates those concerns and proposes the Comparable Decision Record as an operational extension. It contradicts specific source interpretations—weight alone determining priority and paired category swings proving coding error—not the cited literature. It does not empirically confirm the framework’s organizational effects.

[PD] Use the public PDF for the anonymized portfolio narrative. The separate confidential package preserves source locators, raw values, version alternatives, executable calculations and review logs. Do not distribute the private annex as part of the public case.

S1 and internal review. All proposed value calculations remain distinct from historical project outcomes.

REFERENCES / PRIMARY METHODS AND SUBSTANTIVE LITERATURE

Data quality requires context, provenance and a defensible score

These sources support the analytical methods and proposed design. They are not evidence of this project's implementation or results.

R1. Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), 5–33. Fitness for use includes contextual, representational and accessibility dimensions.

R2. W3C (2013). PROV Model Primer. Entities, activities, agents, derivation and versions. Provenance supports assessment of a record; it does not establish its accuracy.

R3. OECD / European Commission JRC (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. Construct definition, normalization, weighting, aggregation and sensitivity.

R4. Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance implications of strategic performance measurement in financial services firms. *Accounting, Organizations and Society*. Measurement systems require evidence linking design and performance.

R5. Campbell, D. T. (1976). Assessing the Impact of Planned Social Change. Original occasional-paper record, ERIC. Evaluative pressure can change an indicator and the process around it.

R6. Goodhart, C. A. E. (1984). Problems of Monetary Management: The UK Experience. *Monetary Theory and Practice*, 91–121. Publisher chapter record; observed relationships may change under control.

R7. 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. Workarounds versus organizational learning.

Original papers, author manuscripts and institutional methods. Citation scope is described in the private methods memo.

REFERENCES / PRIMARY METHODS AND SUBSTANTIVE LITERATURE

Variation, selection and attribution need different tests

These sources support the analytical methods and proposed design. They are not evidence of this project's implementation or results.

R8. NIST/SEMATECH. e-Handbook of Statistical Methods: Individuals Control Charts. Adjacent moving ranges and individuals limits; the case uses the rounded 2.66 coefficient.

R9. Laney, D. B. (2002). Improved Control Charts for Attributes. *Quality Engineering*, 14, 531–537. Extra-binomial attribute dispersion; a count-and-exposure panel is required.

R10. Wheeler, D. J. (2012). *Exact Answers to the Wrong Questions*. SPC Press. Process behavior and statistical interpretation.

R11. The W. Edwards Deming Institute. *The Funnel Experiment*. Institutional explanation of tampering with a stable process.

R12. 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. Selection on extreme measurements and repeated observation.

R13. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. Chance-corrected classification agreement.

R14. Collier, D. (2011). *Understanding Process Tracing*. PS: Political Science & Politics. Necessary and discriminating evidence for specific contribution claims.

Original papers, author manuscripts and institutional methods. Citation scope is described in the private methods memo.

REFERENCES / PRIMARY METHODS AND SUBSTANTIVE LITERATURE

Rollout design must answer the economic decision

These sources support the analytical methods and proposed design. They are not evidence of this project's implementation or results.

R15. Hernán, M. A., & Robins, J. M. Causal Inference: What If. Author-hosted text. Treatment definition, causal graphs, selection and policy estimands.

R16. Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28, 182–191. Cluster rollout structure and power.

R17. Hemming, K., et al. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, 350, h391. Calendar effects and longitudinal cluster implementation.

R18. Hemming, K., et al. (2018). Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ*, 363, k1614. Reporting allocation sequences, periods, implementation and analysis.

R19. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*; author manuscript. Appropriate group-time comparisons for a nonrandom alternative.

R20. Roth, J. (2022). Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*. Nonsignificant pre-trends do not prove identification.

R21. 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. Value of improved choices under information*.

Original papers, author manuscripts and institutional methods. Citation scope is described in the private methods memo.

The revised case supports correction stewardship and a controlled evaluation

[AI] Parallel evidence and editorial work informed the draft. The lead author completed the final review after agent execution stopped. This is internal review, not external peer review.

Major issue	Resolution and remaining boundary
Source and contribution	Verified register cells; original workbooks remain unrecovered. Ownership is registry-attested, with no field-impact credit.
Numerical and source conflicts	Reconciled target-gap denominators, weekly endpoint pairs and source alternatives. Corrected the conditional numerator rounding. No averaging of incompatible versions.
Process signal	Included both individuals and moving-range screens, history-window alternatives and rounding witnesses. Kept the result provisional.
Model and economics	Used an all-eligible-packet loss endpoint, mutually exclusive status categories and explicit hold costs. Cluster/calendar GLS power targets economic clearance; higher variance triggers redesign.
Privacy and comparators	Published only the sorting non-pass index already used in the series. Kept raw levels, exact periods, identities and source paths private.

[AI] Minor issues resolved: time-indexed the DAG; clarified the hypothetical kappa audit size; distinguished weight from normalized priority; corrected non-significance language in falsifiable propositions; and separated reporting loss from customer benefit.

[AI] Referee disposition: suitable as an evidence-reconciliation and operating-design case. Register arithmetic and synthetic models are reproducible. Customer impact, sustained adoption, historical savings and the causal effect of the original tracker remain unestablished.

[AI] Partner critique: the cover gives a COO an actionable decision—close correction dependencies before releasing the weekly decision. The report owner controls version agreement and affected claims; the sponsor commissions the independent value test. A manual workbook edit alone is not the proposed completion criterion.

[PD] Stronger claims require original artifacts, an authoritative metric/target register, recipient and execution records, mature outcomes and a calibrated evaluation.

Final internal review covers executive action, source integrity, model assumptions, arithmetic, public confidentiality and every rendered page. The private package retains calculations and review issues.