

Operational Measurement Assurance

Make the counting rule, population and evidence fit the operating decision

An anonymized operating case on conflicting logistics measurements. The project established a reference canon for table choice, counting unit, filters and interpretation, and documented corrections that changed how reports should be used.

PROJECT CONTEXT / OPERATING ANALYTICS & MEASUREMENT GOVERNANCE

BLUF / BOTTOM LINE UP FRONT

Validate the metric before ranking performance.
Make unresolved evidence a release condition.

ANALYTICAL FINDING

A reported ranking reversal cannot be repaired by applying one common scaling factor. It requires unit-level reconciliation of multiplicity or other rule changes. Even a perfectly closing partition establishes arithmetic consistency, not classification truth.

CONTRIBUTION IN THE SOURCE RECORD

Directed analytical conventions and corrections.
Provisional registry hoop support; query execution and source ownership were shared or external.

EVIDENCE STATUS

Measurement and design case. Conventions are recorded; sustained adoption, causal customer impact and realized savings are unestablished.

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

EXECUTIVE ABSTRACT / DECISION AND OWNERSHIP

Make measurement fitness a condition of operating action

[AI] A technically successful query can answer the wrong operating question. This case converts recurrent discrepancies into a proposed Metric Measurement Contract (MMC): a testable agreement on what a number means and when it may support a decision.

Abstract element	Decision-relevant synthesis
Context [HR]	A distributed delivery network used overlapping tables, row grains and filters. Registered corrections exposed changed rankings, conflicting flags, event-count ratios and incomplete actor evidence.
Objective [PD]	Enable decision owners to distinguish a valid operating comparison from a query that is merely executable or consistent with a reference total.
Approach [HR/PD]	The project documented physical-source, identity, filter and clock conventions. This case reconciles the register and formalizes release tests, escalation and prospective evaluation.
Findings [AI]	Common scaling preserves ranks; reversal needs a unit-level explanation. Partition closure does not validate labels. Dashboard parity can reflect a narrower population rather than corrected truth.
Decision value [PD]	Reduce wrong releases and unnecessary holds together. Evaluate every assigned report task against hidden truth, using loss weights and information-value scenarios kept separate from field outcomes.
Implications [PD]	The COO sponsors the release rule; the data steward owns definitions; source owners resolve lineage; an independent reviewer validates the output and its permitted use.

EXECUTIVE ACTION AND THRESHOLD

[PD] Authorize a bounded MMC pilot. Release a ranking only when its unit, population, identity, time and source-version tests pass. Before scaling, require an independently adjudicated loss reduction whose lower confidence bound exceeds the full incremental cost; retain customer-outcome validation as a separate gate.

Basis: S1 project narrative and metric records. [HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Methods and scope: pp. 10–22.

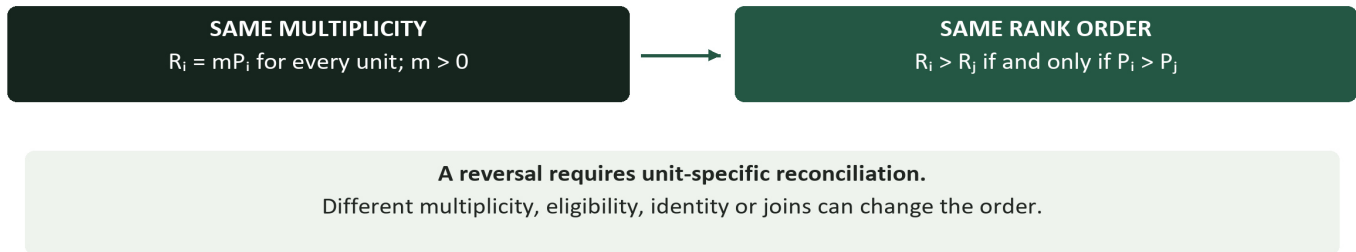
EXHIBIT 1 / GRAIN AND RANKING

A ranking reversal defeats a universal correction factor

[HR] The register reports that replacing row counts with distinct-package counts changed the ordering of defect categories. [AI] The decision consequence is structural: a network-wide scalar cannot repair that ordering.

A COMMON MULTIPLIER PRESERVES THE ORDERING

Analytical proof / the register reports a rank reversal, not a full ranking panel



Source: S1 M08; mathematical inference by Luna Sol. No category effects are estimated.

[AI] Let the row count for category i be $R_i = m_iP_i$, where P_i is its eligible package count and m_i its average multiplicity. If every m_i equals the same positive constant, row and package rankings are identical. A reversal therefore requires differential multiplicity or another simultaneous eligibility, join or identity change.

Required test	Why it changes the decision
Recompute each category	A network average cannot reveal which defect category was incorrectly prioritized. Preserve the row and package rankings and the changes between them.
Freeze the query contract	Hold dates, population, joins and filters fixed while changing the counting unit. Otherwise the source does not isolate grain as the cause.
Validate the key	Distinct identifiers count identifiers. Replans, reused keys or missing links can still misrepresent physical packages.

OPERATING IMPLICATION

[PD] Suspend resource allocation based on the disputed ordering until the unit-level bridge is reproducible. Use COUNT DISTINCT only where distinct-key cardinality matches the estimand; item demand and event workload require different measures.

Source: S1 M08 and recorded grain corrections. Diagram is an analytical proof, not a site dataset. The source supplies a reversal claim, not a complete ranking panel.

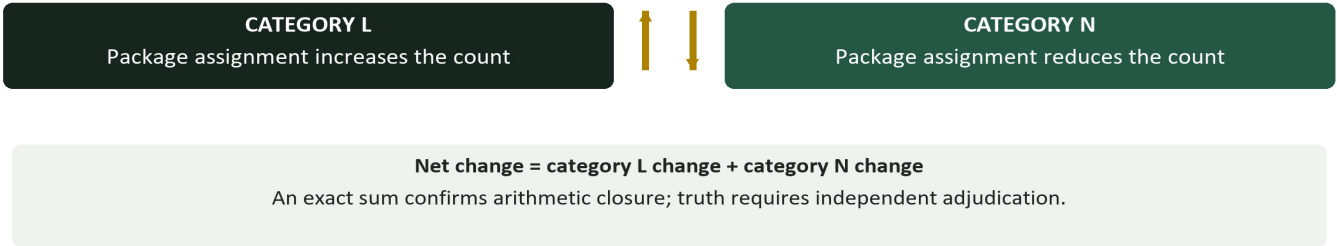
EXHIBIT 2 / PARTITION RECONCILIATION

A closing total can conceal opposite classification errors

[HR] A package-level classification reconciled exactly while row-filtered lost and non-lost categories changed in opposite directions. [AI] The size of a net discrepancy is not a bound on classification error.

NET RECONCILIATION DOES NOT IDENTIFY CLASSIFICATION ERROR

Historical directional bridge / arrows are qualitative, not proportional



Source: S1 M11. Mixed flags and row eligibility must be tested on matching key sets.

[AI] Write the category changes as ΔL and ΔN . The net reconciliation $\Delta T = \Delta L + \Delta N$ can be small when one category rises and another falls. A clean sum follows from an exhaustive, mutually exclusive assignment rule; it does not establish that each package was assigned to its true state.

Check	Evidence needed beyond the total
Logical consistency	Retain all rows for each eligible identity. Test contradictory flags, row eligibility, missingness and the order of filter and aggregation operations.
Semantic correctness	Adjudicate package histories against a separate outcome definition. Taking MAX over a flag implements “ever observed,” not automatically “final true state.”
Overlap discipline	Do not infer an intersection from a separate count of mixed-flag packages. Its eligibility set may differ from the row-filtered comparison.

OPERATING IMPLICATION

[PD] Publish a reconciliation bridge and a disagreement sample. Treat algebraic closure as one acceptance check; require independent label validation before using the partition to attribute failures.

Source: S1 M11. Raw category counts and exact reconciliation remain private. The diagram shows directions only; arrow lengths have no quantitative meaning.

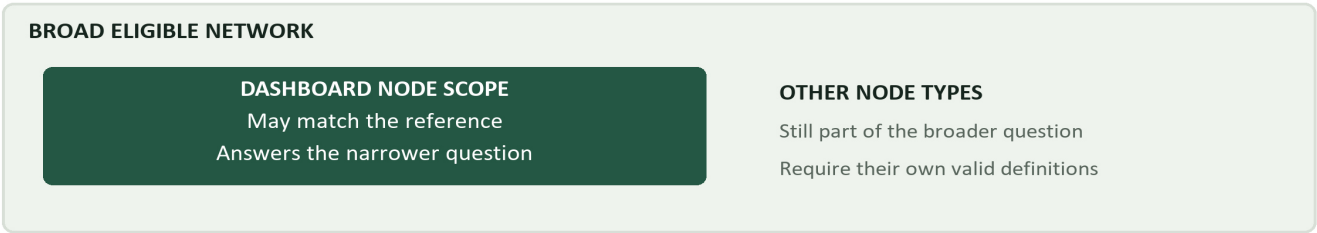
EXHIBIT 3 / POPULATION AND DENOMINATOR

Matching a dashboard can change the question being answered

[HR] Restricting a broad extract to one node type moved its total toward a dashboard and changed the visible leading locations. [AI] Agreement supports that scoped comparison; it does not make the excluded populations irrelevant.

REFERENCE PARITY DEPENDS ON THE CHOSEN POPULATION

Registered scope reconciliation / qualitative set diagram



Neither scope is universal. The sponsor must choose the estimand before the filter.

Source: S1 M06 and population conventions. Areas do not represent relative volumes.

[HR] The source retains an approximate narrative multiplier and a distinct worked calculation. [RC] The worked broad-to-narrow ratio is about 2.57; the narrative says about 2.7. Both support a material scope difference. Neither is a transferable correction factor or a before-and-after performance effect.

Contract field	Publication requirement
Population	Name the business line, node type, country eligibility, active-list version and inclusion dates. Do not let a regional table label define the population.
Numerator / denominator	Reconcile actual key sets, not only their sizes. A numerator universe and a shipment universe can each be plausible yet represent different units.
Cross-group use	Show strata separately until definitions are invariant. Reweight only with explicit eligible volumes; an average of separately rebased indices is not a network rate.

OPERATING IMPLICATION

[PD] Decide whether the sponsor needs dashboard replication or whole-network diagnosis. Preserve both views with their own denominators when both questions matter.

Source: S1 M06 and node/population conventions. Ratios are relative disclosures only; worked and narrative versions are not pooled. Source versions: private evidence audit.

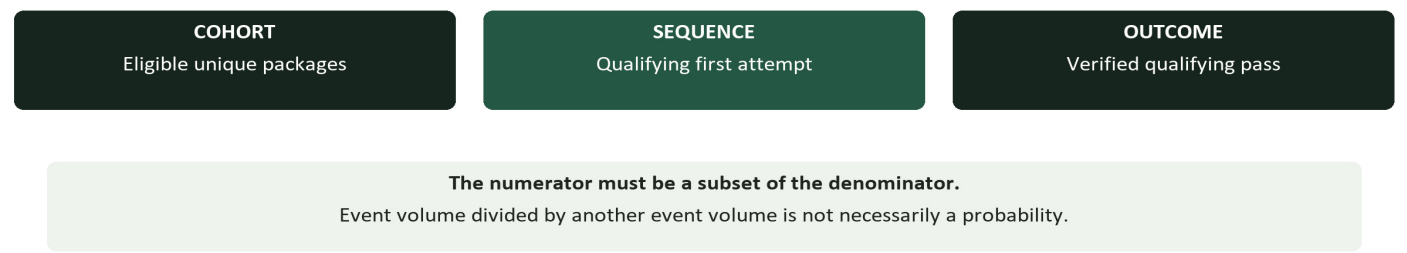
EXHIBIT 4 / EVENT FLOW AND PASS PROBABILITY

Computable event ratios are not automatically valid pass rates

[HR] The source demonstrates that induction and stow event totals can be queried. [AI] A recorded comparison exceeds the range of a probability, showing that event-count availability does not establish a unique-package pass measure.

FOLLOW ELIGIBLE UNITS BEFORE INTERPRETING A PASS RATE

Proposed cohort measurement / event totals can include repeated or unmatched activity



Source: S1 M03–M04; cohort sequence is proposed. No reconstructed internal pass rate is shown.

[AI] A package pass probability requires a numerator that is a subset of the eligible package cohort. Event totals can include repeated scans, different cohorts, late arrivals and cross-window activity. A ratio above one is compatible with those mechanisms; it cannot be interpreted as a probability without a new measurement definition.

Tempting shortcut	Required correction
Divide same-day totals	Follow a fixed package cohort through the relevant event sequence, with defined eligibility, repeat-event treatment and an observation horizon.
Join recent and historical feeds	Reconcile their boundary, deduplication keys and schema. A longer lookback is useful only if the seam creates neither missing nor duplicated events.
Combine available component scores	Verify common unit and date coverage. Component station counts do not identify their intersection or justify complete-case selection.

OPERATING IMPLICATION

[PD] Label an event-throughput ratio as such. Hold any “first-pass” interpretation until the unit-level numerator is nested in its denominator and the mature cohort can be reconstructed.

Source: S1 M03–M04 and component-coverage records. No raw event counts or reconstructed internal rate are disclosed. The range restriction is mathematical, not an estimated effect.

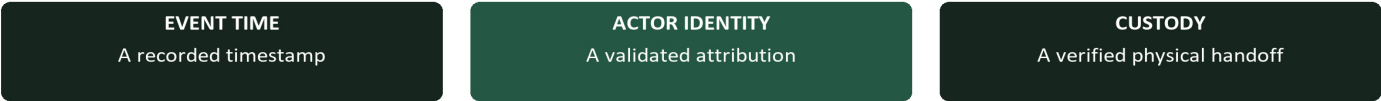
EXHIBIT 5 / CUSTODY AND MISSINGNESS

A timestamp does not identify the actor or establish custody

[HR] Registered fields differ in timestamp availability, actor capture and identity continuity. [AI] These are separate evidence channels; substituting one for another would overstate what the measurement establishes.

OCCURRENCE, ACTOR AND CUSTODY ARE DISTINCT CLAIMS

Analytical evidence ladder / each transition needs additional validation



A populated field does not prove the next claim.
Do not compare capture rates across different eligible populations.

Source: S1 M10–M13, D05/E03. Capture, automation and replans can break the chain.

[AI] Timestamp coverage and actor coverage in the register use different populations. Their ratio is therefore not a same-event capture comparison. Even on a common population, an actor field can reference a system or a joined identifier, while a timestamp can exist without a verified physical handoff.

Inference	Condition required
Recorded event → human action	Validate event generation, actor semantics and automation. Sampling only complete records cannot diagnose missing capture.
Missing actor → no action	Unsupported without a capture model. Missingness may vary by node, workflow, business line or identity replan.
Observed custody share → floor	Requires valid positive links, no relevant false positives, a fixed denominator and one-sided omission. Otherwise use bounds or an unresolved status.

OPERATING IMPLICATION

[PD] Separate event occurrence, actor identity and custody attribution in the report. Resolve unmatched identities and adjudicate exceptions before using the data for individual or partner accountability.

Source: S1 M10–M13, D05/E03 and replan conventions. Cross-population coverage quotients are excluded. Raw identifiers and actor-level evidence remain private.

EXHIBIT 6 / CLOCK AND MATURITY

A median delay does not make a cohort mature

[HR] Category-specific creation lags and local-time corrections changed how operating inputs were aligned with later customer records. [AI] A typical delay is a timing descriptor, not a finality threshold.

DEFINE FINALITY FROM REPORTING COMPLETENESS

Analytical distinction / a median describes timing but not the unresolved tail



A typical delay does not establish that a cohort is complete.
The observation horizon must survive tail, censoring and taxonomy checks.

Source: S1 M05 and clock conventions. No historical completeness threshold is inferred.

[AI] A median identifies the center of a delay distribution. It does not reveal the tail, the share still unreported at a chosen horizon, or whether the observed sample excludes unresolved events. Shifting one series by its median cannot identify the causal lag from operating action to customer outcome.

Risk	Prospective measurement rule
Clock semantics	Retain event time, recorded time, ingestion time and zone conversion separately. An offset histogram is a diagnostic, not proof that every timestamp is correct.
Outcome maturation	Estimate reporting completeness by elapsed time and category from completed cohorts. Freeze a horizon and perform tail/censoring sensitivity before comparison.
Apparent improvement	A coding, deduplication or precision fix changes the measuring instrument. Bridge the definitions and mark the process phase rather than presenting a drop as operational improvement.

OPERATING IMPLICATION

[PD] Attach an as-of date, clock convention and maturity status to each result. Route incomplete cohorts to a provisional view; compare mature like-for-like outcomes before attributing change.

Source: S1 M05 and time/precision corrections. Category medians are cross-sectional summaries, not an ordered process series. Tail probabilities are not available.

EXHIBIT 7 / OPERATING CONTRACT

Turn the reference canon into a versioned release obligation

[HR] The project recorded conventions and corrections; formal adoption and sustained enforcement were not measured. [PD] MMC makes the next step explicit: each decision metric has a named owner, test suite and permitted-use boundary.

MAKE THE METRIC A VERSIONED DECISION CONTRACT

Proposed MMC / every mandatory condition must pass

PURPOSE Question + action	MEANING Unit + population	EVIDENCE Identity + clock	RELEASE Version + owner
-------------------------------------	-------------------------------------	-------------------------------------	-----------------------------------

Release / qualify / hold, with the reason and accountable resolver recorded.
Retest material changes and retain a bridge to affected earlier outputs.

Luna Sol proposed design. Historical documentation is not evidence of enforced adoption.

Owner	Required decision artifact
Business decision owner	Question, action threshold, eligible population and cost of a wrong ranking, unnecessary hold or false release.
Data steward / source owner	Physical source and version, unit key, join cardinality, duplicate policy, taxonomy, clock and maturity definition.
Independent reviewer	Frozen expected outputs, reconciliation tests, disagreement adjudication and a release, qualified use or hold verdict.
Program sponsor	Adoption test, refresh cadence, change log and rollback rules. Separate query-speed evidence from correct decision delivery.

[PD] Every material definition change requires a new contract version, a bridge to affected prior results and an erratum naming the unresolved condition and accountable resolver.

SCALE ONLY AFTER A DECISION-RELEVANT TEST

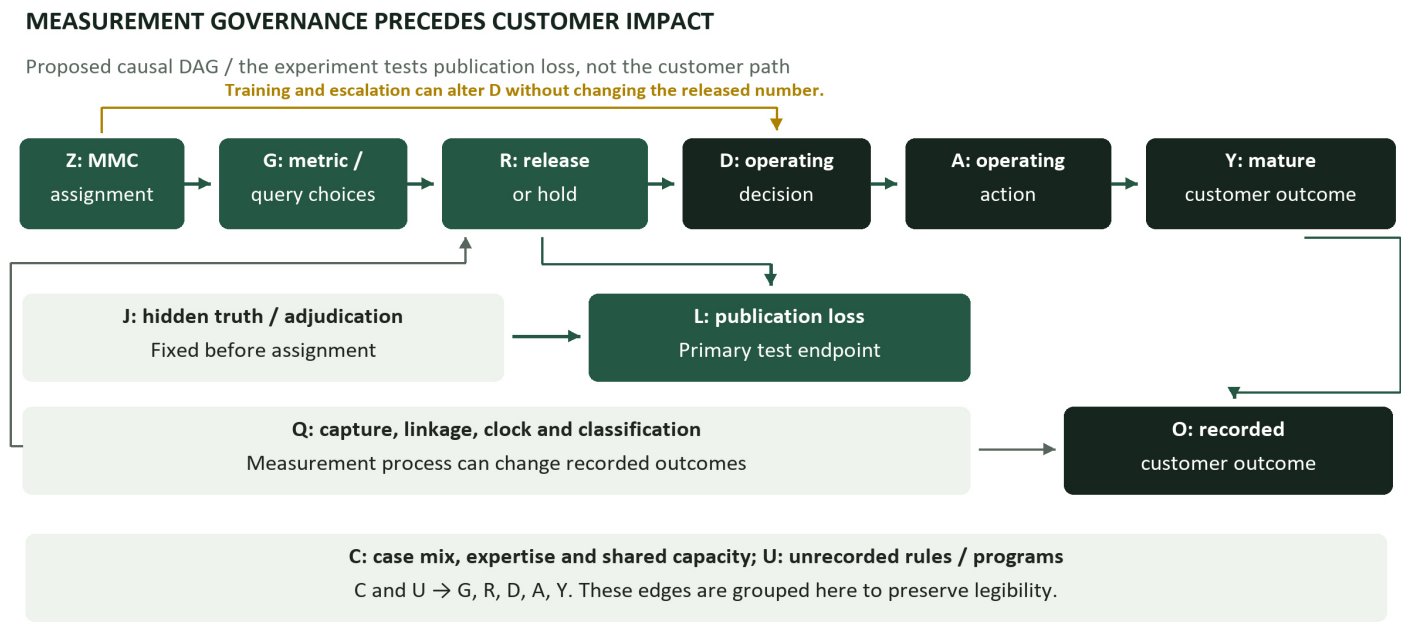
[PD] Compare MMC with usual documentation using complete report-task outcomes, including justified holds and failures to deliver. Independently calibrate the loss weights and economics before claiming value or extending the test to customer outcomes.

Source: S1 governance and ownership narrative. MMC, roles and gates are proposed. Historical conventions do not establish a deployed control system or measured adoption.

TECHNICAL ANALYSIS

A measurement correction changes the evidence before it changes operations

[AI/PD] The proposed Metric Measurement Contract (MMC) binds the decision question to an eligible population, entity grain, source version, transformation, clock and acceptance owner.



Source: proposed model. Residual paths remain unmeasured; assignment to Z does not identify the Y effect.

[PD] Z assigns the MMC package; G represents query, identity and filter choices; R is the released measurement or hold. Independent truth adjudication J and R determine publication loss L. R can influence decision D, action A and mature customer outcome Y. Q captures recording, linkage and clock quality; O is the recorded outcome.

[PD] C names case mix, analyst expertise and shared capacity; U denotes unrecorded rules and concurrent programmes. These can affect measurement and downstream decisions. The proposed benchmark would identify an assignment effect on L, not the effect of A on Y. Training or escalation can also change D directly, so assignment is not automatically an instrument for measurement use.

[AI] The historical record concerns analytical conventions and reported corrections around G and R. A changed R can arise from repaired measurement while Y remains unchanged. Unexplained differences remain compatible with unmeasured capture, scope and operating paths; the graph does not identify which path generated them.

Basis: registered analytical conventions [HR]; proposed MMC graph and benchmark [PD]. No historical customer-effect path is identified.

TECHNICAL ANALYSIS

The counting rule must follow the business entity and question

[HR/AI] The registry describes a package-count correction that changed category ranking. This is a decision-relevant measurement correction, not evidence that fewer packages failed.

Question [PD]	Required unit and operation [PD]
How many packages ever met a condition?	Resolve package identity, freeze the window, aggregate qualifying events, then count eligible packages.
How many events or items occurred?	Count valid events or items using their own keys; repeated package identity can be legitimate.
What was the last valid state?	Apply event order, version and correction rules; an any-event flag does not answer this question.
What cost was incurred?	Sum unique authorized transactions at their financial grain; distinct package counting is not a cost estimator.

[AI] Let category package counts be a and b , and mean row multiplicities be m_a and m_b . If $a < b$, the raw order reverses when $m_a/m_b > b/a$. A common positive multiplier cannot reverse ranking. The reported reversal therefore implicates differential multiplicity or another simultaneous definition change; one universal inflation adjustment is indefensible. Original rows are needed to separate those explanations.

[AI] Filtering event rows before package aggregation can remove a relevant state recorded on another row. Conversely, taking MAX over every historical row can attach a past state to the wrong decision window. Replaying both operators on a discriminating fixture reveals the difference; dashboard agreement alone does not choose the correct construct.

[PD] Version the entity key, alias/replan crosswalk, effective time, duplicate flag and null-key treatment. Test one-to-many joins, legitimate repeated events and corrections explicitly. COUNT(DISTINCT) is a grain-specific implementation choice, not a universal quality rule. SQL execution proves that a query ran; it does not prove that the intended business quantity was measured.

Basis: registered grain correction [HR]; rank-order theorem [AI]; Codd (1970); PostgreSQL aggregate and table-expression semantics. Fixture rules are proposed [PD].

TECHNICAL ANALYSIS

A clean partition cannot authenticate an event or its actor

[AI] Arithmetic closure establishes consistency with a classification rule. It does not establish that the rule describes physical events, authorized custody or final outcomes.

Closure and validity answer different questions

[HR/AI] The register reports a package-level partition after aggregating event flags. A mutually exclusive and exhaustive hierarchy necessarily sums to its chosen population. Conflicting flags can still represent history, correction or an unresolved contradiction. The overlap of separately filtered sets cannot be recovered from a headline count of entities carrying both flags.

[HR/AI] Timestamp and actor evidence were described on different eligible event populations. Their percentages cannot support a same-population coverage quotient. Missing actors do not prove that handling failed to occur; populated actors do not by themselves authenticate a person, physical touch or custody transfer. The registered withdrawal of a custody claim is supported as an analytical decision, not a measured change in frontline behaviour.

Proof requirement [PD]	What must be independently checked [PD]
Event existence	Event meaning, originating service, corrections and duplicate identity.
Actor attribution	Actor type, authorization, identity binding and whether the record was generated automatically.
Event order	Original precision, timezone, daylight-saving rule and valid sequence under replan or replay.
Custody inference	A specified custody transition and source capable of observing it; a generic stage label is insufficient.

[AI] A nonzero UTC/local offset disproves equality of the stored fields, but it does not validate every location-date conversion. Truncating time can merge distinct events and create ties; retaining precision does not establish that clocks were correct. Preserve original timestamps and test known transitions against an independent event source.

[PD] Treat unavailable corroboration as unresolved. Do not convert a filled field, an exact subtotal or agreement with another derivative report into an external truth standard.

Basis: registered provenance and partition findings [HR]; validity critique [AI]; Suriadi et al. (2017); Fellegi and Sunter (1969); W3C PROV-DM.

TECHNICAL ANALYSIS

A reported lower bound requires a fixed denominator and valid links

[AI] Missing linkage can make a count incomplete without making every reported ratio a lower bound. Direction depends on the numerator, denominator and classification errors.

State the assumptions that make a floor valid

[AI] Let p be the share of a complete fixed eligible population with the specified true event. Let v be the share with a verified positive and u the unresolved share. With error-free positive and negative classifications, p lies in $[v, v+u]$. This is a logical identification interval, not a confidence interval. If false-positive mass is bounded by e , the lower endpoint becomes $\max(0, v-e)$; without a bound it can be zero.

[AI] The interval requires resolved entity identities, no false links in the verified class and nonnegative unit contributions. If the denominator is itself restricted to successfully linked records, the observed share need not be below the full-population share. Negative financial adjustments also invalidate a simple positive-count-to-value floor. The source’s floor convention therefore remains conditional until these assumptions are tested.

Missingness mechanism [PD]	Required reporting treatment [PD]
Unlinked or replanned identity	Retain the original eligible entity; report possible duplicate or missed-link bounds.
Missing date or actor	Record unknown status; do not silently drop it or substitute a measured zero.
Unobserved source partition	Name the coverage boundary; an accessible subset is not the full population.

[HR/AI] The registered median delay measures creation timing among recorded cases in a particular extraction. A median does not identify the tail or demonstrate cohort completion; right truncation can change it. A shift in recognition date is not automatically the maturation of the original customer cohort.

[PD] Freeze the original eligibility date and observation horizon. Audit capture by age, source and business line; retain unresolved outcomes and report worst-case bounds. Reconcile competing rosters and exposure sources as separate specification branches before choosing an accountable reference.

Basis: registered identity, missing-field and maturity limitations [HR]; binary bounds derived here [AI]; Rubin (1976); Beresteanu and Manski (2000).

TECHNICAL ANALYSIS

Validate the metric contract against independent cases

[PD] A reproducible metric specification must identify what is counted, which population is exposed, when evidence becomes eligible and who accepts exceptions.

Metric family	Validity contract and invariance test [PD]
Package failure risk	Unique upstream eligible units; at least one prespecified event by a fixed horizon; no outcome-dependent denominator.
Event-flow ratio	Aligned event populations and clocks; do not call repeated-event counts a unique-package first-pass probability.
Coverage or provenance	State the eligible event, field semantics and capture path; compare like populations and retain unknowns.
Reported performance series	Freeze scope, node/roster version, numerator, exposure unit, maturity, sign and comparison baseline.
Publication loss	Independently frozen task truth, original deadline, all assignments and nonoverlapping loss categories.

[HR/AI] Matching a narrower dashboard scope shows conditional reproduction, not total-network truth. Differing channel and node populations can require legitimate exclusions. A missing exposure crosswalk or unvalidated retention seam remains an unresolved construct issue even when the resulting figure looks plausible.

[PD] Before the trial, draw a random core of 200 future action-context or metric-specification cases from a declared frame; this is not a claim about available historical cases. Two blinded coders independently classify grain, linkage, cohort, maturity and release eligibility against a frozen codebook. Adjudicate using original evidence, preserving both initial labels.

[PD] Under independent cases, 200 observations give a worst-case agreement margin of about ± 6.9 percentage points at a normal 95% screen. Shared source families require a design-effect correction. Kappa precision needs prevalence-specific pilot simulation; report the confusion matrix, raw agreement, class prevalence and interval alongside kappa. Supplement rare exceptions, but do not treat their unweighted sample as representative.

[AI] Agreement establishes reproducibility of classification, not validity. A common mistaken canon can produce perfect agreement; independent counterexamples must remain capable of overturning it.

Basis: proposed measurement audit [PD]; registered scope and seam conflicts [HR]; Wang and Strong (1996); Cohen (1960); original SQL semantics.

TECHNICAL ANALYSIS

A restated ranking is not a process signal or treatment effect

[AI] Definition changes, source substitutions and corrected duplicates can move a metric without changing operations. Variation analysis starts only after the measurement contract is stable.

Separate operating time from source and specification versions

[PD] For commensurate ordered observations, compute $MR_t = |x_t - x_{t-1}|$, individual limits $\text{mean}(x) \pm 2.66 \times \text{mean}(MR)$, and moving-range upper limit $3.267 \times \text{mean}(MR)$, the convention used privately. Retain timing, phase and denominator. Short or degenerate sequences provide tentative screens, not calibrated proof of stability. An out-of-limit point triggers investigation, not an identified cause.

[HR/RC/AI] The private inventory contains three screens: an ordered series of three quoted weekly changes and two mechanical nonprocess comparisons. All quoted changes lie within their tentative individual limits; source-vintage invariance is unconfirmed. There is no commensurate operating panel. Query forms, lag conventions and source corrections must not be pooled into an intervention series; absent observations are never interpolated.

[PD] Rate charts need repeated compatible numerator/exposure pairs. Laney p-prime additionally requires a defensible subgroup/dispersion model; large row counts alone do not justify it. An unknown capture process cannot be repaired by widening statistical limits. Within-limit movement is not established improvement or deterioration.

Selection and business context change the interpretation

[AI/PD] No fixed worst-site cohort with a comparable untreated follow-up is supplied. As a hypothetical sensitivity, selecting the highest tenth of standard-normal baseline errors gives a mean of 1.755 SD. With assumed repeat reliability 0.3, 0.6 or 0.9, expected untreated reversion is approximately 1.23, 0.70 or 0.18 SD. These are assumptions, not estimated shares of this correction.

[PD] Freeze focus/non-focus membership before follow-up. Estimate effects within business-line, node and source-family strata; test interactions with uncertainty. A dedupe rule useful for repeated package records should not suppress legitimate repeated events, and an exclusion valid for one partner population should not erase another population's customer risk.

Basis: source-comparison limitations [HR/AI]; proposed SPC and selection sensitivity [PD]; NIST individuals charts; Deming; Barnett et al. (2005).

TECHNICAL ANALYSIS

Attribute the convention separately from the underlying data estate

[HR/AI] The register supports a lead role in analytical conventions and question-to-source mapping. It does not establish sole execution of every query, upstream repairs or realized operating benefits.

Contribution claim	Process-tracing assessment [HR/AI]
Directed the question-to-source map and population convention	Provisional hoop: explicit registry attribution is consistent with direction; original authorship and timing require authentication.
Established counting and provenance rules	Provisional hoop for analytical ownership; field-level tests were also attributed to analysts or agents.
Measured query-performance improvement	Assigned to a collaborator in the record; no sole-lead measurement claim.
Built the source tables or published data estate	Unestablished and outside the attributed role; source teams retain ownership.
Secured adoption, upstream fixes or customer savings	Unestablished; no instrumented adoption panel or causal outcome comparison is supplied.

[AI] Multiple derivative summaries are not independent corroboration. A registered personal statement or session activity is not a smoking gun for human execution. No doubly decisive contribution test is established. The excluded unrelated application remains outside this case even if a search returned a similar topic.

Consolidated limitations and reproducibility

[HR/AI] Source integrity: the selected register is derivative; reported live tests do not replace reopened query text, result hashes or original logs. Measurement: identity, population, retention, mature outcome and source-to-published-series mappings retain unresolved branches. Attribution: lead direction, collaborator execution and platform ownership are distinct. Identification: no contemporaneous causal comparison identifies customer or cost impact.

[AI/PD] Transport: a constructed publication benchmark does not establish production error prevalence or field adoption. Economics: scenario costs are assumptions; raw financial-table totals are not benefits. Privacy: raw counts, field populations and identity crosswalks remain private, including companion-case inversion checks. Reproduction requires frozen source locators, transformations, truth fixtures, query versions and calculations. The absence of a recovered completion trace does not establish whether further work occurred elsewhere.

Basis: source role and limitation records [HR]; process-tracing assessment [AI]; original evidence recovery status governs any later upgrade.

TECHNICAL ANALYSIS

Test the publication contract before claiming operating impact

[PD] Randomize independent analyst/source-family domains to MMC availability and orientation or usual documentation. Test correctness under the same fixed evidence; do not execute customer-facing actions.

Prospective intention-to-treat and independent truth

[PD] Each domain receives a frozen bank of 24 tasks: 18 have a defensible publishable result and six require a hold because a mandatory source condition is unresolved. These are synthetic design counts, not historical workload. Truth, population, correct grain, deadline and permissible exceptions are fixed before assignment. The intervention cannot relabel a task to earn success.

[PD] The primary estimand is $\Delta = E[L(\text{usual})] - E[L(\text{MMC})]$, mean cost-weighted publication loss per originally assigned task at its original 48-hour deadline. Positive Δ is benefit. All tasks remain in the denominator; late or missing submissions receive the preregistered hold classification. Wrong rates/ranks, unnecessary holds, justified holds and correctly supported publications have separate prepriced categories. Normalization uses one fixed usual-loss reference.

[PD] Blinded adjudicators apply an independently constructed truth manifest, including entity-replan and conflicting-flag counterexamples. Where truth itself cannot be resolved, preregister an uncertainty/hold rule rather than silently dropping the task. Audit unsupported releases, excess holds and completion separately; economic success cannot waive mandatory evidence standards.

[PD] Cluster at the unit sharing analyst learning and source rules; split or merged teams must not create artificial independent observations. The finite planning model includes shared domain difficulty, changes in variance with effect size, exact null/break-even calibration and sensitivity. The design must detect the same-horizon economic threshold; higher task counts cannot substitute for independent domains.

[AI/PD] Parallel allocation limits learning carryover; shared staff still require isolation. A later authorized field rollout could stagger adoption across source/operating clusters and estimate group-time ATT against not-yet-treated clusters, following Callaway–Sant’Anna. Require common support, no anticipation, conditional parallel trends and interference checks. Inspect baseline balance and event-time plots with joint and equivalence diagnostics; nonsignificant pre-trends are not proof. Freeze customer eligibility and maturation, then collect count/covariance inputs and power that separate trial. The benchmark does not identify execution effects or mature customer benefit.

Basis: proposed domain-randomized publication benchmark [PD]; synthetic model assumptions and economic threshold are disclosed on the model pages. No benchmark has been executed.

TECHNICAL ANALYSIS

Four propositions make the measurement contract falsifiable

[AI/PD] MMC connects relational identity, evidence quality and decision release. The case demonstrates specification failures; the framework’s prospective average benefit remains to be tested.

Proposition and mechanism [PD]	Disconfirming observation or boundary [PD]
P1: explicit grain and identity rules reduce incorrect category ranking on mixed-grain tasks.	No prespecified reduction on fresh independently constructed tasks; a faulty identity crosswalk can reverse the expected benefit.
P2: provenance gates reduce unsupported actor or custody attribution.	Unsupported releases persist or increase after unknowns remain in the original denominator.
P3: cohort, clock and denominator contracts reduce specification-induced disagreement.	Common-contract replays retain disagreement beyond justified uncertainty; a shared wrong definition can still agree.
P4: versioned corrections and accepted-owner routing reduce recurrence of known measurement defects.	The same defect reappears at unchanged or higher frequency in new eligible outputs despite confirmed rule availability.

[AI] Codd grounds relational representation; Fellegi and Sunter distinguish links from identity truth. MMC turns those concerns into a publication gate. The rank-order proof diagnoses differential distortion, not business impact. Rubin and event-log research preserve uncertainty when capture and event meaning are unresolved; the custody retraction does not prove absent handling.

[AI] Deming’s tampering argument presumes a meaningful measure; MMC first checks its definition. Campbell and Goodhart motivate testing whether pressure to hit a reported target changes classification or selection. This is a prospective risk, not evidence of deliberate manipulation. Ittner and colleagues caution against equating measurement sophistication with performance. Tucker and Edmondson distinguish local workaround from upstream learning; documented rules do not establish source repair.

[AI/PD] Composite-indicator design makes weighting and compensation explicit. MMC treats mandatory identity, population and authority gaps as gates: a high score elsewhere cannot compensate for a missing prerequisite. Do not average separately rebased group indices into a network rate; restore compatible quantities and declared exposure weights. The OECD handbook informs this design choice without certifying the local metric.

[PD] Preregister material effect and guardrail thresholds. A lower confidence bound above break-even supports the modeled investment gate; an upper bound below it contradicts the economic benefit claim. An interval spanning the threshold remains inconclusive. Neither statistical success nor arithmetic closure overrides an unresolved mandatory source condition.

Basis: proposed MMC propositions [PD]; Codd; Fellegi and Sunter; Rubin; Suriadi et al.; Deming; Campbell; Goodhart; Ittner et al.; Tucker and Edmondson; OECD. No imported effects.

PROSPECTIVE MODEL / PUBLICATION ASSURANCE

Test the released measurement, not the number of rules recorded

[PD] Randomize independent analyst/source-family domains to the Metric Measurement Contract or usual documentation. All values below are hypothetical benchmark assumptions.

[PD] Freeze 24 report tasks per domain before assignment: 18 contain sufficient evidence for a correct publication; six require a hold because a mandatory source condition is unresolved. Distinct task instances follow a balanced truth manifest. Teams use only one protocol; isolate staff and source-family dependencies. No production query, ranking or release is performed.

Terminal state at 48 hours	Loss units	Why it matters
Feasible, correct publication	2	Ordinary output effort
Feasible, withheld or unsubmitted	6	Ordinary effort plus delay burden
Materially wrong rate or ranking released	10	Ordinary effort plus correction burden
Unpublishable, justified hold	2	Valid diagnosis and routing effort

[PD] Primary ITT benefit is usual-minus-contract mean loss across every assigned report at the original deadline. Classify each report once; multiple defects do not earn multiple penalties. Grain, eligible population, denominator, maturity and ranking/tie rules are preregistered. Late, missing and withheld outputs remain in the fixed denominator.

A HOLD IS NOT AUTOMATIC SUCCESS

[RC/PD] All-hold loss is 5.00 per task, above the usual-law 4.31. Contract-law loss is 3.36. The loss function values correct publication and justified restraint; it cannot be improved by suppressing every feasible output.

[PD/RC] Proposed controlled benchmark, not observed productivity or savings. Loss weights are assumed resource proxies; false release and feasible-output coverage also require unweighted guardrails.

PROSPECTIVE MODEL / PUBLICATION ASSURANCE

Shared source difficulty changes precision before it changes the average

[PD] A source-family mistake can contaminate many reports. Twenty-four correlated reports are not twenty-four independent domains.

Conditional truth class	Usual verdict probabilities	Contract verdict probabilities
Publication feasible	Correct 0.60; wrong 0.25; hold 0.15	Correct 0.72; wrong 0.12; hold 0.16
Publication not feasible	Wrong release 0.18; justified hold 0.82	Wrong release 0.08; justified hold 0.92

[PD] Equally likely domain types shift wrong-release probability by plus or minus 0.05, taking the same mass from correct publication or justified hold. Conditional on type, tasks are independent. Average the domain laws only after composing all tasks. Arms and domains are independent; source instances and adjudication are hidden during tuning.

[PD] Lambda interpolates the two verdict laws. Higher lambda stochastically lowers loss within each truth class; the variance is recalculated along this path. Exact independent-arm convolution supplies the null and economic-boundary laws. The model assumes known probabilities; pilot uncertainty requires recalibration before a real study.

Required validity control	Implementation before the benchmark [PD]
Reference truth	Independent dual coding, declared arithmetic tolerances and tie handling; adjudicate disagreements without treatment labels. Audit source completeness and oracle errors.
Safety and transport	Preregister false-release and feasible-output coverage margins and separately power them. The fixture mix does not establish production prevalence, error costs or customer effects.

NONNEGOTIABLE BOUNDARY

[PD] A reproducible answer can still be wrong if the reference population or truth manifest is wrong. Agreement, exact finite-law power and a low mean loss do not certify measurement validity.

[PD] Benchmark truth is independently constructed, not assumed infallible. One domain is one randomization unit; no paired-trial assumption or pipeline-count substitution is used.

PROSPECTIVE MODEL / PUBLICATION ASSURANCE

A publication-assurance trial must clear the resource threshold

[PD] Require 80% power to detect break-even against zero and 80% probability that the one-sided 97.5% lower bound exceeds break-even at the full planning alternative.

[RC/PD] Usual loss is indexed to 100; modeled reduction is 22.04 points. Net resource value = reduction × avoidable share a × non-overlap h – incremental cost 12. Base a = h = 1 gives break-even 12; this is not a cash-saving forecast.

Domains per arm / total	80% MDE	Economic power	Both gates
24 / 48	14.62	50.4%	Fail
48 / 96	10.33	80.7%	Conditional
80 / 160	8.05	95.6%	Conditional

[RC/PD] Minimum: 48 per arm under the base law. A 50% larger domain shift requires 64; candidate 80 per arm clears that stress with 88.5% economic power. Recruitment feasibility is unknown. With a = 0.75 and h = 0.90, 254 per arm are required. At a = 0.50, even non-overlap h = 1 leaves the full effect below break-even.

Full-effect net / a	h = 1.00	h = 0.90	h = 0.80
0.50	-0.98	-2.08	-3.18
0.75	4.53	2.88	1.23
1.00	10.04	7.84	5.63

[PD] Cost 12 allocates incremental canon checks 5, maintenance 4 and support 3 to the same service pool; ordinary task effort is already in the terminal losses. Research adjudication is separate. Exact two-sided alpha = 0.05 and boundary false clearance at most 2.5% apply only within the known nuisance family.

[PD] Separate field value = eligible volume × mature customer ITT reduction × avoidable cost/outcome × non-overlap – incremental cost. Population and maturity must match; unavailable inputs leave field value unestimated and unpowered.

[PD/RC] Finite-law inversion; effect-dependent variance. Shares uniformly scale net terminal burden; otherwise reprice categories. Calibrate weights, oracle reliability and dependence before field use.

PROSPECTIVE MODEL / PUBLICATION ASSURANCE

Buy evidence that changes the repair decision before releasing a metric

[PD] Information value is a separate planning question: whether further evidence changes the choice of routine preparation, grain repair or source-scope validation.

[PD] States are clean for the declared tests, grain-biased, or scope-mismatched, with priors 0.50 / 0.30 / 0.20. Scope mismatch takes priority when both defects exist. Other defect classes require a larger state space. Every final release still requires the mandatory measurement contract.

Planning action / state	Clean	Grain bias	Scope mismatch
Prepare routine publication	12	-30	-40
Fund grain repair	6	18	-20
Fund source-scope validation	-3	-4	9

[RC/PD] Prior-optimal grain repair is worth 4.40; perfect information adds at most 8.80. The three-label signal is correct with probability 0.45, 0.70 or 0.90; wrong labels split the remainder equally. Optimize the planning action after each signal.

Signal quality	EVSI	Evidence cost	Net EVSI
Low	0.0000	0.40	-0.4000
Base	2.3950	1.20	1.1950
High	6.4450	2.50	3.9450

[PD] Utilities already include the downstream planning-action cost; information-acquisition cost is additional. These utilities do not use the benchmark loss improvement and must not be added to it. A favorable evidence value supports investigation, not an unsupported metric release or a booked operating benefit.

WHAT THIS ANALYSIS CHANGES

[AI/PD] The contract makes publication conditional on the unit, population, denominator and source evidence. It preserves a valid hold while making unnecessary abstention visible. Historical adoption, prevented error cost and mature customer impact remain unestablished.

[PD/RC] Finite-state Bayesian model with hypothetical priors and net utilities. Evidence can change a decision while remaining insufficient to authorize publication.

Reproduce the measurement logic without exposing the operating estate

[AI] This is a measurement and operating-design case. Registered corrections support a decision narrative; a proposed evaluation tests whether formalizing those rules creates incremental value.

Claim class	Meaning and boundary
[HR] Historical record	Registered observation, convention, correction or attributed role. Derivative source summaries are distinguished from authenticated original artifacts.
[RC] Recalculated	Reconciliation, normalized comparisons and hypothetical planning results recomputed from declared inputs. Correct arithmetic does not authenticate a source.
[AI] Analytical inference	Rank-order theorem, partition and missingness limits, or a bounded interpretation of the record. These are not observed average treatment effects.
[PD] Proposed design	MMC, future evidence collection, tests, release gates and economic scenarios. None is presented as an executed experiment or realized benefit.

Reproduction sequence

[PD] Verify the frozen workbook and extract hashes. Reconcile the grain, scope, partition, timing and provenance branches against exact source cells. Preserve conflicts as separate specifications. Recompute the synthetic model and decision-information scenarios, independently verify their distributions, and regenerate every exhibit and PDF.

[PD] The separate private package contains raw values, source identities, locators, recovery results, claim ledger, calculation code, assumptions and review dispositions. It excludes the full original workbook and production credentials. Reproducing the planning model requires no operating-system access or external data pull.

WHAT THE NEW ANALYSIS CHANGES

[AI] A universal multiplier cannot repair reversed rankings, and a closing total cannot authenticate a classification. These findings change the acceptance test: validate the unit-level decision evidence rather than merely reconciling the reported number.

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

RESEARCH REFERENCES / METHODS AND INTERPRETATION

Ground counting, linkage and event meaning in explicit definitions

Primary research and official methods inform the analysis and proposed test. They do not independently corroborate the historical project.

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *CACM*, 13(6), 377–387. Original IBM publication record. Relational representation and consistency motivate explicit identity and grain; the paper does not validate this project's keys.

PostgreSQL Global Development Group. PostgreSQL 18, §9.21 Aggregate Functions. Official documentation distinguishes row counts, non-null expression counts and aggregate behavior. This is a semantics reference; internal-engine behavior still requires verification.

PostgreSQL Global Development Group. PostgreSQL 18, §7.2 Table Expressions. Official documentation distinguishes joins, row filtering and grouping. The case's counterexample explains why changing their order can change the question.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3), 581–592. Original article. Ignoring missingness requires assumptions about the observation mechanism; this case establishes none of those assumptions for actor or linkage capture.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *JASA*, 64(328), 1183–1210. Original authorized reprint. False links, missed links and undecided cases are distinct. The proposed identity audit does not presume that a matching string establishes custody.

Suriadi, S., Andrews, R., ter Hofstede, A. H. M., & Wynn, M. T. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information Systems*, 64, 132–150. Publisher excerpts. Imperfect event records need diagnosis before process interpretation.

Moreau, L., & Missier, P., eds. (2013). PROV-DM: The PROV Data Model. W3C Recommendation. Entities, activities, agents and derivations inform the versioned evidence ledger. Lineage does not itself certify accuracy.

Research verified in this pass or identified as prior-series verification in the private research memo. Publisher metadata and abstract-only access are distinguished from full-text review.

RESEARCH REFERENCES / METHODS AND INTERPRETATION

Separate statistical uncertainty from unidentified quantities

Primary research and official methods inform the analysis and proposed test. They do not independently corroborate the historical project.

Beresteanu, A., & Manski, C. F. (2000). Bounds for MatLab: Draft Version 1.0. Author methods manuscript. Incomplete observations can imply intervals rather than point identification. The binary bounds in this case are derived directly; the software package was not used.

NIST/SEMATECH. e-Handbook of Statistical Methods, §6.3.2.2, Individuals Control Charts. Official formulas use successive moving ranges. Short sequences and measurement-definition changes here prevent calibrated claims of operating stability or improvement.

Laney, D. B. (2002). Improved Control Charts for Attributes. *Quality Engineering*, 14(4), 531–537. Publisher abstract and metadata. Dispersion matters for attribute charts; a large denominator does not turn heterogeneous or changing definitions into a valid rate series.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20, 37–46. Publisher metadata. Kappa supplements the proposed confusion matrix and raw agreement; agreement alone is not measurement validity.

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. *IJE*, 34, 215–220. Primary abstract. The worst-cohort sensitivity is an illustrative model, not a measured component of this project's results.

Collier, D. (2011). Understanding Process Tracing. *PS: Political Science & Politics*, 44, 823–830. Primary publication. Diagnostic evidence tests constrain personal contribution claims; a derivative attribution does not establish sole execution or ownership of the source estate.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225, 200–230. Author manuscript. Relevant to a later staggered field rollout; it is not the estimator for the proposed randomized publication benchmark.

Research verified in this pass or identified as prior-series verification in the private research memo. Publisher metadata and abstract-only access are distinguished from full-text review.

RESEARCH REFERENCES / METHODS AND INTERPRETATION

Connect measurement governance to learning and economic decisions

Primary research and official methods inform the analysis and proposed test. They do not independently corroborate the historical project.

The W. Edwards Deming Institute. The Funnel Experiment. Official exposition. Adjustments responding to routine variation can worsen a process. MMC adds a prior check that the measurement definition is stable; historical tampering is not established.

Campbell, D. T. (1976). Assessing the Impact of Planned Social Change. Original institutional record. Consequential indicators can distort the activity they represent; dashboard parity and audit passes should not replace independent decision validity.

Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance Implications of Strategic Performance Measurement in Financial Services Firms. AOS, 28, 715–741. Author manuscript. Measurement sophistication is not evidence of an operating benefit in this case.

Tucker, A. L., & Edmondson, A. C. (2003). Why Hospitals Don't Learn from Failures. CMR, 45(2), 55–72. Author working paper. Distinguishes local workarounds from system-level learning; documented analytical corrections do not establish upstream repair.

OECD / European Commission JRC (2008). Handbook on Constructing Composite Indicators. Official handbook. Normalization, weights and sensitivity matter; missing mandatory evidence cannot be compensated by unrelated high component scores.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. JMIS, 12(4), 5–33. Original scan. Contextual fitness goes beyond a populated field or an exact sum.

Strong, M., Oakley, J. E., Brennan, A., & Breeze, P. (2015). Estimating the Expected Value of Sample Information Using the Probabilistic Sensitivity Analysis Sample. MDM, 35, 570–583. Author repository. The case uses its own finite-state information model, not the paper's regression estimator.

Goodhart, C. A. E. Problems of Monetary Management: The U.K. Experience. 1975 conference; revised volume listed by the RBA as 1976. Publisher archive documents the source of Goodhart's law. No monetary-policy result or historical gaming diagnosis is imported.

Research verified in this pass or identified as prior-series verification in the private research memo. Publisher metadata and abstract-only access are distinguished from full-text review.

FINAL REFEREE REPORT / INTERNAL AGENT-ASSISTED REVIEW

Approve the case as measurement design, with an explicit impact boundary

[AI] This is an author-led review assisted by evidence, methods and numerical agents. It is not journal peer review, client endorsement or certification of a firm ranking.

Major issue	Disposition in the revised case
Rank and partition claims	Category reversal is distinguished from site-scope ranking. Common-scaling impossibility is proved; exact partition closure is not treated as true classification.
Measurement and provenance	Cross-population capture quotients, probability interpretations of event ratios and unconditional custody floors are removed. Cohort, identity and maturation assumptions are explicit.
Evidence and contribution	Registered direction is separated from collaborator execution and platform ownership. Recovery results and exact source locators are retained privately; causal impact remains unidentified.
Evaluation and value	The all-task publication-loss test includes holds and failures. Hypothetical loss weights, dependence and economic thresholds are disclosed; information value is separate from operating impact.
Reproducibility and privacy	Source arithmetic, model outputs and original diagrams are checked. Raw values and identifying evidence remain private. Actual page inspection and file hashes bind the delivery review.

Partner critique: would a COO act on page one?

[AI] Yes: require a defensible metric before assigning an operating response, and authorize a bounded contract pilot with a data steward and independent reviewer. The cover makes the decision, new finding and personal-contribution boundary visible without implying a realized saving.

Minor revisions and remaining limitations

[AI] Review corrected the object of the ranking reversal and source identifiers, narrowed actor and maturity claims, and prevented abstention alone from satisfying the synthetic utility model. The consolidated technical limitations govern historical attribution and field transfer.

REVIEW DISPOSITION — PASS WITHIN STATED SCOPE

[AI] The case supports measurement forensics, a new interpretation of reported corrections and a reproducible proposed evaluation. It establishes no historical average customer effect, adoption effect or realized financial benefit.

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