The Measurement Trap Designing What Counts

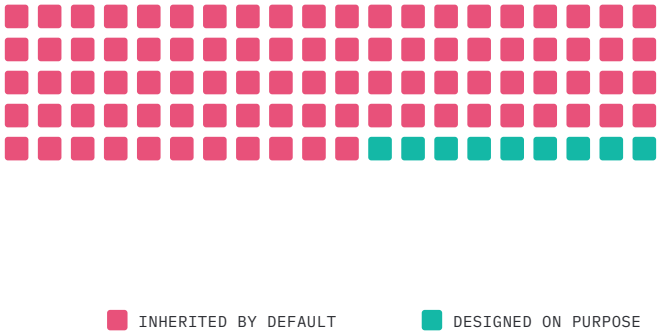
James E. Coleman

STRATEGIC DESIGN · WHITE PAPER SERIES · 2025

ABSTRACT *What you measure reshapes what you get. Metric design is the silent architecture beneath the three failures this series has already traced — the intelligence gap, the valley of death, the speed-trust paradox. Yet most measurement systems are inherited by default, not designed with intent.*

FIG. 1 — 100 NUMBERS A COMPANY TRACKS INHERITED: 91%

Picture the 100 numbers a company watches. **More than 90 of them** were never actually chosen — they were copied from the past, not designed for the goal. Only a handful were picked on purpose.



WE'VE KNOWN FOR 70 YEARS — AND IGNORED IT

1956

Ridgway publishes the **first peer-reviewed account** of what happens when a number becomes the goal.[2]

1975

Goodhart's Law and Kerr's "folly of rewarding A while hoping for B" are **both formalized**.[3]

5

distinct **failure modes** of metric fixation have been named since — the ways a measure quietly corrupts the thing it was meant to track (§02).

>90%

And yet today, more than nine in ten of the metrics a company tracks were **inherited by default** — never actually designed for the goal.[1]

§ 01 Goodhart's Law in Practice

The measure *becomes the target.*

The economist Charles Goodhart observed, in a 1975 critique of monetary policy, that any statistical regularity a government leans on for control tends to collapse once the pressure is applied — a relationship that held while it was merely observed breaks the moment it is made a lever.[2] Two decades later the anthropologist Marilyn Strathern distilled the idea into the form most people now quote: when a measure becomes a target, it ceases to be a good measure.[4] The compression is exact. A metric is a proxy — a visible stand-in for something valuable but hard to see directly. Attach consequences to the proxy and people optimize the proxy, not the thing it was meant to represent.

The pattern is not confined to economics. Donald Campbell, surveying social reform, arrived at the same law from the opposite direction: the more any quantitative indicator is used for decision-mak-

ing, the more it will be subject to corruption pressures and the more apt it will be to distort the processes it was intended to monitor.[5] Test scores drive teaching to the test. Arrest counts reshape what police choose to see. Hospital waiting-time targets rearrange the queue rather than shorten it. In each case the number improves while the underlying reality it was supposed to track quietly diverges.



A vernier caliper — what you can measure finely depends entirely on choosing the right instrument.

Photo: Auckland Museum · CC BY 4.0

> *"A metric is a hypothesis about what matters, frozen into a number. The danger is not that the hypothesis is wrong — it is that, once it starts to steer behavior, no one remembers it was ever a hypothesis at all."*
>
> — JAMES E. COLEMAN, METRIC ARCHITECTURE FRAMEWORK

This is the thread running beneath the three prior papers in this series. WP 1.0's intelligence gap is, at root, a decision steered by the wrong indicators. WP 2.0's valley of death is sustained by publication and grant metrics that reward novelty over translation. WP 3.0's speed-trust paradox turns on trust being hard to measure and therefore easy to ignore. Each failure has a measurement system underneath it, inherited rather than designed. Goodhart's Law is not a curiosity at the edge of these problems — it is the mechanism at their center.

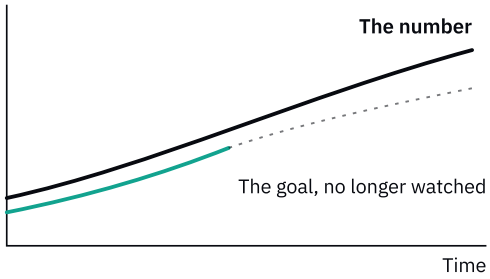
What gets measured *gets gamed.*

The failure is not random. It follows recognizable modes, documented across six decades. The earliest controlled account is Ridgway's 1956 survey of performance measurement, which showed that single, composite, and multiple measures alike produced dysfunctional side effects — narrowed focus, inter-unit conflict, and the suppression of unmeasured work.[1] Two decades later Steven Kerr catalogued the organizational version in his classic essay on "the folly of rewarding A, while hoping for B" — reward systems that pay for one behavior while the designer sincerely hopes for another, and are then surprised to get the behavior they paid for.[3]

A subtler mode is *surrogation*: the cognitive substitution of the measure for the thing it represents, so that managers come to believe the proxy *is* the goal. Controlled accounting experiments show the effect strengthens precisely when compensation is tied to the measure — the incentive that was meant to focus attention instead collapses the strategy into its indicator.[6] Once surrogation sets in, gaming no longer feels like cheating; it feels like doing the job.

FIG. 2 — FIVE WAYS THE NUMBER LEAVES THE THING IT MEASURED

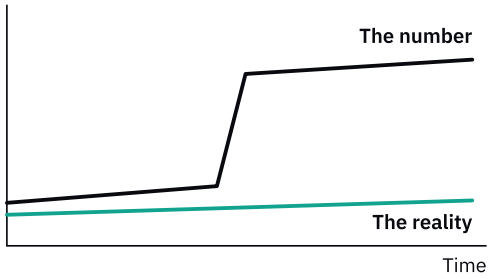
STATE 1 OF 5



THE PROXY REPLACES THE GOAL

Surrogation

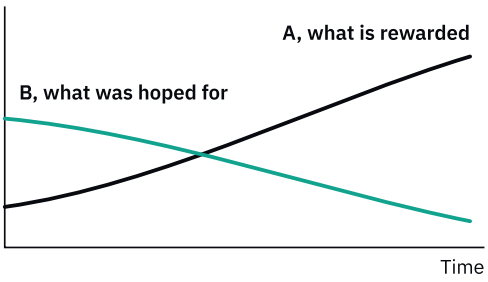
The proxy is mistaken for the goal. Teams optimize the indicator as though it were the objective itself, losing the strategy it was meant to represent.[6]



THE NUMBER MOVES, THE WORK DOES NOT

Gaming

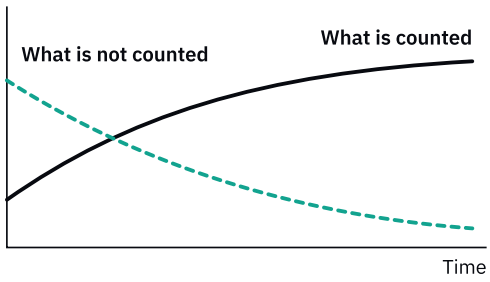
The number is moved without moving the reality — reclassifying cases, timing work to the measurement window, shifting effort to what is counted at the expense of what is not.



A IS REWARDED, B IS DROPPED

Goal displacement

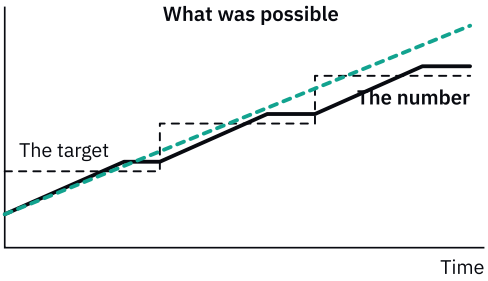
Hitting the metric displaces the mission. Reward for A while hoping for B, and B is quietly abandoned in favor of whatever A measures.[3]



ONLY THE COUNTABLE GETS EFFORT

Measuring the measurable

Effort concentrates on what is easy to quantify; the important-but-unmeasured — judgment, relationships, long-horizon value — is neglected because it does not appear on the dashboard.[1]



EFFORT STOPS AT THE LINE

Ratchet and threshold effects

Targets create distortions at their edges — bunching just above a cutoff, sandbagging to keep next period's quota low, and effort that stops the instant the threshold is crossed.

A diagram of the relation, not a chart of data: the shapes are drawn from the descriptions and no values are plotted. Black is the number on the dashboard. Green is what it was meant to track.

Each mode shares a root cause: the measurement system treats a proxy as if it were costless and complete. It is neither. Every metric imposes a tax — the attention it diverts, the behavior it bends, the territory it renders invisible by not counting it. A system that accounts only for the benefits of measurement and never for its costs will reliably over-measure the measurable and under-serve the mission.

§ 03 What Resists Measurement

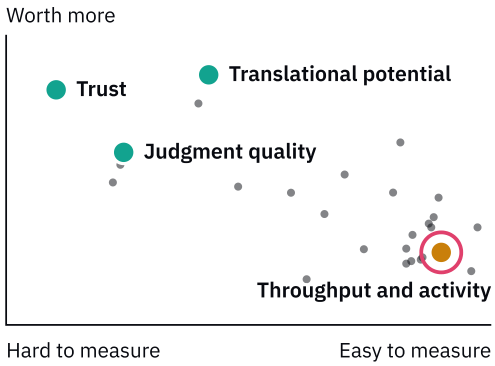
The assets that matter most *quantify the least.*

The historian Jerry Muller, in his study of the modern vogue for metrics, names the recurring error "metric fixation" — the conviction that the only knowledge worth having is quantifiable, that such metrics should drive reward, and that to do otherwise is to indulge in guesswork.[7] The trouble is that the assets this series has argued are decisive are exactly the ones that resist clean quantification.

Judgment quality, translational potential, and trust are high-value and low-measurability at once — and that combination is precisely what makes them so easy to starve.

The reason is structural, not a failure of effort. Trust, as WP 3.0 argued, accrues slowly through consistency across contexts and reveals itself mainly in its absence; it has no natural unit. The quality of a decision is only legible long after it is made, and often confounded by luck. A discovery's translational potential is a probability distribution over futures, not a figure. Where no honest direct measure exists, organizations reach for a proxy — citations for research impact, response time for service quality, headcount for capability — and the proxy, under pressure, drifts from the thing it was standing in for. The danger is not that these assets go unmeasured; it is that a convenient proxy gets optimized in their place.

FIG. 3 — WHAT IS WORTH MOST IS HARDEST TO MEASURE STATE 4 OF 4



HIGH VALUE, VERY HARD TO MEASURE

Trust

High value, very low measurability. No natural unit; visible mainly when it fails. Common proxy — transaction volume or survey score — rewards the appearance of trust over its substance.

HIGH VALUE, HARD TO MEASURE

Judgment quality

High value, low measurability. Legible only in hindsight and confounded by luck. Common proxy — short-run outcomes — punishes good decisions that drew bad draws and rewards the reverse.

HIGH VALUE, HARD TO MEASURE

Translational potential

High value, low measurability. A distribution over uncertain futures, not a figure. Common proxy — citations or near-term grants — selects for novelty and fundability, not for the discoveries most worth carrying across the valley of death.

LOWER VALUE, VERY EASY TO MEASURE

Throughput and activity

Lower strategic value, very high measurability. Counts, latencies, volumes. Cheap and precise — and therefore over-represented on every dashboard, crowding out the assets that actually compound.

Positions are ordinal, taken from the descriptions (very low, low, very high). Nothing here was measured. Each small dot is a unit of dashboard attention, and it drifts toward whatever is cheapest to count.

Aggregation adds a second, quieter failure: a single number can improve while the people behind it are harmed. A population-level average — a readmission rate, screening coverage, a patient-satisfaction score — can rise even as a subgroup's outcomes fall, because the gains of the majority mask the losses of a minority. An organization optimizing the headline figure is then, in effect, optimizing against the populations its average conceals. Measuring what counts therefore means disaggregating it — tracking outcomes across the groups the aggregate would otherwise average away, so that disparity is visible in the measure rather than hidden by it.

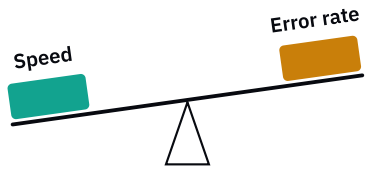
The honest response is not to abandon measurement for the hard-to-measure, nor to pretend a proxy is the real thing. It is to hold the proxy loosely — to name it as a proxy, to watch for its drift, and to keep human judgment in the loop precisely where the number is weakest. That discipline is the subject of the final section.

§ 04 Designing Honest Metrics

Metrics are infrastructure — *so design them.*

If measurement systems are inherited by default, the corrective is to treat them as what they are: designed systems, in the sense this series has used the term throughout. A metric is not a neutral readout of reality; it is an intervention that reshapes the reality it reads. Designed well, it directs attention toward what matters. Designed by accident, it directs attention toward whatever happens to be countable. The difference is not sophistication of instrumentation — it is intent.

FIG. 4 — FOUR WAYS TO DESIGN A METRIC THAT STAYS HONEST STATE 1 OF 4

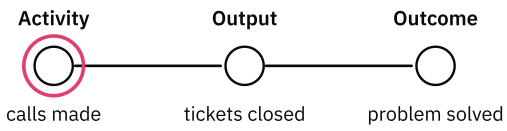


Push one down and the other comes up

TWO MEASURES THAT PULL AGAINST EACH OTHER

Pair every metric

No metric travels alone. Pair each with a counter-metric that captures what optimizing it would damage — speed against error rate, volume against quality, output against burnout. The pair makes gaming visible, because moving one at the other's expense shows up immediately.

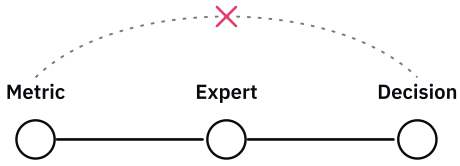


Move the measure to what the work is for

FROM BUSYNESS TO RESULT

Measure outcomes, not activity

Count what the work is *for*, not how busy it looks. Activity metrics — calls made, tickets closed, papers published — are easy to game and weakly tied to value. Outcome metrics are harder to move dishonestly.

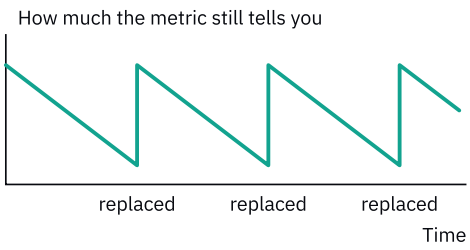


The number informs a person. It does not decide.

A NUMBER INFORMS A JUDGMENT

Keep humans in evaluation

Where an asset resists measurement, do not hand the judgment to a proxy. Use metrics to inform expert evaluation, not to replace it — explicitly for trust, judgment, and translational potential, the assets §03 showed are least quantifiable.



EVERY MEASURE WEARS OUT

Revise as they are gamed

Goodhart's Law is a property of time, not of bad metrics. Any measure decays once it becomes a target, so treat metrics as perishable: monitor them for drift, retire them when captured, and rotate them before the gaming hardens into culture.[2]

Four diagrams of relations, not charts of data. Choose a principle to hold it.

> *"Intelligence, translation, and trust are accumulated assets — and so is a measurement system's honesty. It compounds when metrics are designed with intent and maintained against gaming, and it depreciates the moment a proxy is allowed to quietly become the goal."*
>
> — JAMES E. COLEMAN, WP 6.0

The series thesis holds here as everywhere: durable advantage comes from architecture, not from more resources. More measurement is not better measurement. An organization that pairs its metrics, measures outcomes, keeps judgment in the loop, and revises under pressure will out-decide one with ten times the dashboards and no design behind them. What you choose to count is the quietest and most powerful design decision you make — because, in the end, you will become what you measure.

↳ The series thruline

The gap between knowledge and action is an infrastructure problem, not a resource problem — crossed by systems designed around the people who do the crossing. Here the gap runs from what is measured to what matters. The same move recurs across the series: build the system the people who do the crossing currently lack.

§ R References — Primary Sources

Sources *& notes.*

Every empirical and historical claim in this paper is cited to a primary or seminal source below. The one quantitative figure without a primary source — that more than 90% of tracked metrics are inherited rather than deliberately designed — is labeled a modeled estimate on the page; it is illustrative of the paper's thesis, not a measured value.

1. [1]Ridgway VF. “Dysfunctional Consequences of Performance Measurements.” *Administrative Science Quarterly* 1956; 1(2): 240–247. The first controlled account of how single, composite, and multiple performance measures each generate dysfunctional side effects. doi:10.2307/2390989.

2. [2]Goodhart CAE. “Problems of Monetary Management: The UK Experience.” In *Monetary Theory and Practice: The UK Experience*, pp. 91–121. London: Macmillan, 1984 (originally presented 1975). The origin of “Goodhart's Law”: a statistical regularity breaks down once relied upon for control. doi:10.1007/978-1-349-17295-5_4.

3. [3]Kerr S. “On the Folly of Rewarding A, While Hoping for B.” *Academy of Management Journal* 1975; 18(4): 769–783. Classic analysis of reward systems that pay for one behavior while hoping for another. doi:10.2307/255378.

4. [4]Strathern M. “‘Improving ratings’: audit in the British University system.” *European Review* 1997; 5(3): 305–321. Source of the widely quoted formulation: “when a measure becomes a target, it ceases to be a good measure.” doi:10.1017/S1062798700002660.

5. [5]Campbell DT. “Assessing the Impact of Planned Social Change.” *Evaluation and Program Planning* 1979; 2(1): 67–90. “Campbell's Law”: the more a quantitative indicator is used for decision-making, the more it corrupts and distorts the process it monitors. doi:10.1016/0149-7189(79)90048-X.

6. [6]Choi J, Hecht GW, Tayler WB. “Lost in Translation: The Effects of Incentive Compensation on Strategy Surrogation.” *The Accounting Review* 2012; 87(4): 1135–1163. Experimental evidence that tying pay to a measure causes managers to substitute the measure for the strategy it represents. doi:10.2308/accr-10273.

7. [7]Muller JZ. *The Tyranny of Metrics*. Princeton, NJ: Princeton University Press, 2018. Survey of “metric fixation” across medicine, education, policing, and business, and of the recurring costs of mistaking the measurable for the valuable. doi:10.1515/9780691191263.

SERIES CLOSING STATEMENT

"Intelligence and trust are both accumulated assets. They compound when invested consistently and depreciate rapidly when neglected. The organizations that will define the next era of economic activity are those that treat both as core infrastructure — not as emergent properties of scale, but as designed systems, maintained with intention, evaluated against outcomes."