Goodhart's Law and the Metrics We Choose

S. LINDQVIST, 14 DECEMBER, *for* PHILOSOPHY OF MEASUREMENT

When a measure becomes a target, it ceases to be a good measure. This is Goodhart's Law, and it is usually cited as a warning about bureaucratic incentive design (Strathern 1997). I argue here that the law is not merely a caution about specific metrics gone wrong, but a structural claim about any system where an agent both optimises for a measure and controls the process the measure observes — and that this structural reading explains why the failure recurs across domains that share almost nothing else in common.

1 The standard reading

The standard reading of Goodhart's Law treats it as a story about gaming: a call-centre worker rewarded for call volume rushes calls; a school rewarded for test scores teaches to the test; a hospital rewarded for short waiting times reclassifies patients rather than treating them faster. In each case, the metric was a reasonable *proxy* for the underlying goal until it became the target itself, at which point the agent's optimisation pressure found the cheapest path to a high score rather than the intended path to the underlying goal.

This reading is correct as far as it goes, but it treats each instance as a separate failure of measurement design — a particular metric that turned out, in hindsight, to be gameable. That framing invites a particular kind of response: design a better metric, close the loophole, add a second metric to catch what the first one missed. I want to suggest this response mistakes the depth of the problem.

2 A structural claim

Consider what all three examples above share, beneath their surface differences. In each, the agent being measured has some control over the *process* that produces the outcome the metric observes, and that control is exercised with knowledge of the metric. This is not a fact about call centres, schools or hospitals specifically — it is a fact about any principal-agent relationship where the agent has more granular knowledge of the process than the principal has of the outcome.

Under this reading, adding a second metric does not close the gap; it relocates it. If a school is now measured on test scores *and* teacher-reported student wellbeing, the optimisation pressure simply moves to whichever of the two measures has more slack in it — and since the agent, not the principal, has the finer-grained knowledge of where that slack is, the principal cannot know in advance which measure will absorb the pressure. This matches an observation from complex-systems theory: a control system with fewer degrees of freedom

than the system it is trying to control cannot, even in principle, capture that system's full behaviour (Ashby 1956). Two metrics do not have more degrees of freedom than the one process generating both, if the agent understands the relationship between them.

3 Why this matters for design

If the structural reading is right, the practical implication is not "design better metrics" but "reduce the gap between what is measured and what the agent actually controls, or reduce how much the agent knows about what is measured." Both are harder and less comfortable prescriptions than adding a second KPI, which may explain why organisations reliably choose the second KPI instead: it produces a defensible-looking response to a failure without requiring anyone to give up the legibility that having a single clean number provides.

This also predicts something the gaming-story version does not: that Goodhart failures should recur even in organisations that have already been burned by one and have explicitly tried to correct for it, because the correction (a second or third metric) does not close the structural gap the argument above identifies. Anecdotally, this matches what is actually observed — organisations rarely escape metric-gaming by adding metrics, and more often escape it (when they do) by reducing an agent's discretion over the process itself, which is a much more disruptive change than adjusting a scorecard.

4 Conclusion

Goodhart's Law is usually taught as a lesson about picking better proxies. I have argued it is better understood as a claim about the limits of any measurement relationship in which the measured party has more knowledge of the underlying process than the measuring party has of the outcome. This reframing does not make the problem easier to solve, but it does explain why the easy solution — adding another metric — keeps failing to solve it.

References

Ashby, W. Ross (1956). *An Introduction to Cybernetics*. Chapman & Hall.

Strathern, Marilyn (1997). "'Improving ratings': audit in the British University system". In: *European Review* 5.3, pp. 305–321.