# Designing and evaluating metrics

Sean J. Taylor Follow
May 24 · 9 min read ★

Now more than ever, we need clear thinking about what measurements we should use to understand our world, our products, and ourselves. A **metric** is simultaneously 1) a designed artifact, 2) a lens through which we observe phenomena, and 3) way we set and monitor goals. The goal of this post is to articulate how I think about metrics (after spending several years on dozens of data science projects). I propose five key properties of metrics that represent key tradeoffs in the design process, as well as a depiction of the lifecycle of a metric.

that "everything measured..." quote

I've thought about metrics a lot because I believe measurement forms the foundation of science, as well as being a key technology for improving policy and business outcomes. There are countless examples: better timekeeping technology enabled humans to travel farther and map entire continents, systematic collection of astronomical observations revolutionized astronomy, and making maps of cholera cases enabled John Snow to determine whether water sources were causing disease. Investments in our ability to capture data and measure outcomes often precede step-function changes in our understanding of the world and the ability to better solve problems.

① capture data
② measure outcome

difficult thing abt measurement is: we dont know what to measure. bad choice → eg. Goodhart's law

Our collective investment in measurement creates our distributed perceptual system, focusing our attention on chosen *properties* of a chosen set of *events*, on chosen *scales*. Metrics become part of the language we use to discuss what's happening, framing how we make decisions by denominating our goals, problems, and constraints.

sometimes we think of solutions as goals. impact: it doesnt solve the relevant issue.

Perhaps most importantly, goal metrics become the target of coordinated processes to optimize (or more likely satisfice) within organizations. Just as we should be careful what we wish for, *we should be careful what we optimize for*. We should devote substantial effort to choosing the topology and y-scale of the hill we're climbing implied by our metrics, and ensure that we encode our risks and downsides as counter-metrics we can monitor alongside our goals.

## Five properties of metrics

In this section I discuss five main properties to keep in mind when designing a metric. Improving these properties leads to natural tradeoffs you face as you develop a measurement strategy that helps meaningfully improve a product or user experience. But the properties extend broadly beyond data science for business into many other scientific areas. You'll notice an emphasis on statistical and causal properties here, because my experience is based heavily in improving products through experimentation.

### Cost

I start with cost because it is the most neglected aspect of measurement. You can (basically) measure anything if you are willing to pay an arbitrarily large cost. Cost can entail money, calendar time, employee time, user time (interrupting users to ask them stuff), computation, or technical debt. Metric costs often imply important tradeoffs. I have noticed a gradual trend toward using human labels, survey responses, or external data sets as metric strategies, all of which introduce substantial complexity, delay, and error into measurements.

Though we often take cost as fixed or as a binding constraint, it's important to point out that in many cases we can trade time, money, or effort for better measurement. This tradeoff is challenging to manage because we must also estimate the payoff from having better metrics, and how that may propagate into downstream product or decision quality.

## Simplicity

Metrics are designed artifacts and people have a preference for simple artifacts. The worst possible metric is one that people mistrust, second-guess, or ignore. I have noticed that metrics can often be improved through normalization (which tends to focus them) and made worse through combination (which tends to unfocus them). For instance in sports analytics, we find it exceedingly useful to adjust successful outcomes by dividing by opportunities for success (e.g. batting averages) or accounting for context (e.g. home field advantage). But we do not try to make a batting average also capture how often a hitter hits home runs.

dont force it to do > 1 thing at once.

It's worth emphasizing that when normalizing metrics, finding appropriate denominators can be extremely challenging.

In past projects I have tried to stretch the limits of metric simplicity with what I would call "modeled metrics" which are the outputs statistical models which smooth and improve the precision of estimates. I have not yet seen one of these approaches fully succeed — simplicity can be sacrificed, but it must produce some commensurate improvement in other properties.

## Faithfulness

There are an unfortunately large number of ways your measurement can fail to accurately represent the concept you care about. Two of the most important ones I have observed in practice are metrics without **construct validity** or that have some kind of **sampling bias**. Metrics without construct validity measure the wrong thing. Measures with sampling bias measure it for the wrong set of units (e.g. people, items, events, etc).

We often sacrifice construct validity in order to gain simplicity or low cost, and I have often seen teams gradually add complexity or invest time and effort to improve it. A common struggle with construct validity is in using human-labeled data — interpretation of labeling guidelines can vary and produce labels that mean different things to different people.

Increasingly, products will incorporate user feedback or labeled data in metrics (surveys, bug reports, crowdsourced labels), introducing problematic sampling bias. How can we ever know if people giving feedback will be representative of who we care about? If we can't sample randomly (as is the case with surveys or content ratings), we can never completely solve this problem and must accept it as irreducible source of error in our metrics. It's worth pointing out that even simple metrics like counting "likes" on a social media app may suffer from a large bias in participation rates, and likely to reflect the behaviors of a biased subset of users.

Two interesting examples of failures of metric faithfulness:

1. There is evidence that clicks on display ads are not predictive of sales. If you used clicks as a metric for your ad campaign, you would optimize for an irrelevant outcome; clickers do not resemble buyers.
2. There is very low correlation between sentiment measured using text of social media posts and self-reported emotional states based on surveys. If you measure people's happiness from their Twitter or Facebook posts, there's a good chance you're getting it wrong. prob bias → ppl who feel certain emotions may not tweet/post at all perhaps

## Precision

Precision is the simplest of the five considerations — more precision is better and noisy metrics mean we can't separate the signal from the noise. That means we can't say for sure if a change is due to something we caused or not (experimentation), and we can't understand if a change is occurring over time (trends and anomalies). Three things are useful to know about precision:

1. You can gain substantial precision through transformations of metrics, either through taking logs, winsorizing, or even fancier techniques.
2. Normalizations can substantially improve precision of a metric. If the numerator is very skewed and the denominator is as well, then the ratio creates a much lower variance metric.
3. Summing or average several metrics is useful for precision. If you have a few relatively uncorrelated ways of measuring the same thing, their sum will be less noisy. The price is reduced simplicity, and perhaps less causal proximity (next section).

Often there's an inherent tradeoff between precision and faithfulness. Although they are what we ultimately care about, metrics that capture financial outcomes (sales, revenue, or profit) can be quite noisy because of skewed distributions. Counting discrete outcomes like transactions or unique customers (which are binarizing a continuous outcome) will have bounded variance.

### Causal proximity

A good metric can be affected by causes under your control. Deng and Shi (2016) define a property called sensitivity, which is composed of precision (above section) and typical effect sizes. I think separating these two properties is interesting and I use "proximity" to capture the idea that a metric "close" in "causal space" (e.g. in a path along a causal DAG) to policies you are capable of changing.

When causal proximity is low, you will not often move the metric with your product changes because a sequence of outcomes must occur for you to cause an effect. Low causal proximity is why it is wildly ineffective to use use profit or revenue as a metric for most product changes. We must choose a metric with higher proximity and rely on a theory about how that is useful for some ultimate goal — a sacrifice of faithfulness.

bc it might take a while for the effect of the changes to take effect. → low causal proximity

↳ trade off. prox faithfulness is not straightforward

We sometimes call this strategy a *proxy metric*, acknowledging that it may not be exactly what we care about, but that is a concept on which we can detect effects. For longer-term outcomes of interest, there is exciting and recent work on *surrogate indices*— estimates of the (more faithful) long term outcome estimated from short-term metrics.

→ purpose is to detect effects

Very high causal proximity is not always desirable! Metrics that move too readily are trivially game-able and better serve as monitoring metrics (e.g. to detect some the adverse effect of introducing a bug) or for verifying that an experiment is doing what you expect it to (i.e. manipulation checks).

need to separate monitoring metrics & metrics that we optimize towards

## Dignity

*(Just kidding.)*

## Lifecycle of a metric

In my experience, metric design is iterative and involves the cooperation of multiple stakeholders through a lengthy process where many steps are repeated. This picture is an idealized version of that process. You can see it’s actually a bunch of nested loops that you can keep going around ad infinitum. That’s because metric design is never really done. Metrics, just like code, are evolving artifacts that need to be tested, re-evaluated, tweaked, and eventually replaced when they no longer satisfy criteria of users.

Some specific thoughts about the phases:

- **Discussion**: It's a good idea to formalize the process of choosing metrics and formalize their evaluation by gathering requirements. At the risk of sounding repetitive, we are *designing an artifact* that will be used by many people, we need to carefully understand their varied needs and manage tradeoffs between them. Many metrics are chosen because they are 1) convenient or 2) low cost, but going cheap here may severely limit your ability to learn later.
- **Validation**: I have been surprised at what tends to convince people to believe new metrics: a small number of examples that agree with their intuitions tends to be persuasive. Showing that they move in expected directions when good or bad product changes have been introduced is a good story telling that will help people build trust. Deng and Shi argue for having a corpus of known good/bad experiments to evaluate that metrics move in the expected direction, which I think is a nice luxury if you have run many historical experiments.
- **Experiment**: I have noticed many teams fail to get well-powered experimental estimates for the metrics they care about most. I worked on a product at Facebook where we ran many experiments for several months with no substantive effects because our metrics were noisy and had low causal proximity. If you can't cause a (statistically and practically) significant effect for your metric, then it is not very useful. You may need to sacrifice some faithfulness for causal proximity or precision, or be willing to pay a higher cost. Bad metrics should not even be included in your experiment analyses or as part of your experimentation platform — they lower the signal/noise ratio of experimental results!

question! is it possible to understand the tradeoff b/w key metrics early - even b4 we decide on them as metrics?

this is so u can quickly figure out whether - it's working as expected or not asap

- **Optimize**: What happens after we optimize a metric? We might think this is a fantasy and we can always do better, but for many metrics there is a point of saturation, or a point at which it starts to erode something else we care about. A central challenge at many businesses is understanding tradeoffs between key metrics and making principled decisions to manage those tradeoffs effectively. It's worth noting that after attempts at optimization, metrics may become less effective at

not always up

can saturate

erode other metrics

capturing what they intend, a phenomenon known as Goodhart’s Law.

## Thanks

This post was inspired by a lot of conversations and collaborations with friends and former co-workers: Tom Cunningham, Eytan Bakshy, Annie Franco, Amaç Herdağdelen, and George Berry.

Data Science Metrics Statistics Experimentation