7

Metrics for Experimentation and the Overall Evaluation Criterion

> Tell me how you measure me, and I will tell you how I will behave
> – *Eliyahu M. Goldratt (1990)*

> The first rule is that a measurement—any measurement—is better than none. But a genuinely effective indicator will cover the output of the work unit, and not simply the activity involved. Obviously, you measure a salesman by the orders he gets (output), not by the calls he makes (activity)
> – *Andrew S. Grove in High Output Management (1995)*

Why you care: *To design and run a good online controlled experiment, you need metrics that meet certain characteristics. They must be measurable in the short term (experiment duration) and computable, as well as sufficiently sensitive and timely to be useful for experimentation. If you use multiple metrics to measure success for an experiment, ideally you may want to combine them into an Overall Evaluation Criterion (OEC), which is believed to causally impact long-term objectives. It often requires multiple iterations to adjust and refine the OEC, but as the quotation above, by Eliyahu Goldratt, highlights, it provides a clear alignment mechanism to the organization.*

From Business Metrics to Metrics Appropriate for Experimentation

As discussed in Chapter 6, data-driven organizations often use goal, driver, and guardrail metrics to align and execute on business goals with transparency

and accountability. However, these business metrics may not be directly useful for online experimentation, as metrics for experimentation must be:

- **Measurable:** Even in an online world, not all effects are easily measurable. For example, post-purchase satisfaction can be challenging to measure.
- **Attributable:** To compute the metrics for experiment purposes, we must be able to attribute metric values to the experiment variant. For example, to analyze whether the Treatment is causing a higher app crash rate than the Control, we must be able to attribute an app crash to its variant. This attribution may not be available for metrics provided by other data providers, such as third parties.
- **Sensitive and timely:** Experiment metrics must be sensitive enough to detect changes that matter in a timely fashion. Sensitivity depends on the statistical variance of the underlying metric, the effect size (the delta between Treatment and Control in an experiment), and the number of randomization units (such as users). As an extreme example of an insensitive metric, you could run a controlled experiment and look at the stock price of the company. Because the ability of routine product changes to impact the stock price during the experiment period is practically zero, the stock-price metric will not be sufficiently sensitive. At the other extreme, you could measure the existence of the new feature (is it showing?), and that will be very sensitive, but not informative about its actual value to users. Between the two extremes, click-throughs on the new feature will be sensitive but highly localized: a click-through metric will not capture the impact on the rest of the page and possible cannibalization of other features. A whole-page click-through metric (especially if penalized for quick-backs where users come back quickly), a measure of "success" (like a purchase), and time-to-success are usually good key metrics sensitive enough for experimentation. See Dmitriev and Wu (2016) for an in-depth discussion on sensitivity. Here are a couple more common examples:
  - With ads revenue, it is common for a few outliers to have a disproportionally high influence on revenue, like clicks with very high cost-per-click. While a dollar is a dollar and these expensive clicks should be included in business reporting, these large outliers inflate variance and make it harder to detect Treatment effects. For this reason, you could consider a truncated version of revenue for experiments as an additional more sensitive metric (see Chapter 22).
  - Consider a subscription contract that has a yearly renewal cycle. Unless you are willing to run a year-long experiment, it will be hard to measure

> the impact on the renewal rate. For this case, instead of using renewal rate in experiments, it is common to find surrogate metrics, such as usage, which are early indicators of satisfaction that will lead to renewals.

Based on these considerations, you can see that not all metrics that are used for business reporting purposes are appropriate for experimentation. We do agree with Andrew Grove's quotation above: when in doubt, measure more, but more importantly: think hard about what you are optimizing for. Declaring time-on-site as a metric to optimize without qualifiers like (good/successful session) will lead to interstitial pages and a slow site, which will increase the metric in the short term, but cause abandonment in the long term.

In general, for experimentation, you will be choosing the subset of business goal, driver, and organizational guardrail metrics that meet these measurability, computability, sensitivity, and timeliness characteristics. Then you may need to further augment that metric set with:

- Additional surrogate metrics for your business goals and drivers
- More granular metrics, such as feature-level metrics to help understand movements of specific features. For example, a page-click-through rate may be broken into click-through rate on the dozens of features on the page.
- Additional trustworthiness guardrails (see Chapter 21) and data quality metrics
- Diagnostic and debug metrics that provide information too detailed to track on an ongoing basis but useful when drilling into a situation where the goal, driver, or guardrail metrics indicate a problem.

Given all the different taxonomies and use cases for metrics, a typical experiment scorecard will have a few key metrics, and hundreds to thousands of other metrics, all of which can be segmented by dimensions, such as browsers and markets.

Combining Key Metrics into an OEC

Given the common situation where you have multiple goal and driver metrics, what do you do? Do you need to choose just one metric, or do you keep more than one? Do you combine them all into single combination metric?

While some books advocate focusing on just one metric (*Lean Analytics* (Croll and Yoskovitz 2013) suggest the One Metric that Matters (OMTM) and *The 4 Disciplines of Execution* (McChesney, Covey and Huling 2012) suggest focusing on Wildly Important Goal (WIG)), we find that motivating but an

oversimplification. Except for trivial scenarios, there is usually no single metric that captures what a business is optimizing for. Kaplan and Norton (1996) give a good example: imagine entering a modern jet airplane. Is there a single metric that you should put on the pilot's dashboard? Airspeed? Altitude? Remaining fuel? You know the pilot must have access to these metrics and more. When you have an online business, you will have several key goal and driver metrics, typically measuring user engagement (e.g., active days, sessions-per-user, clicks- per-user) and monetary value (e.g., revenue-per-user). There is usually no simple single metric to optimize for.

In practice, many organizations examine multiple key metrics, and have a mental model of the tradeoffs they are willing to accept when they see any particular combination. For example, they may have a good idea about how much they are willing to lose (churn) users if the remaining users increase their engagement and revenue to more than compensate. Other organizations that prioritize growth may not be willing to accept a similar tradeoff.

Oftentimes, there is a mental model of the tradeoffs, and devising a single metric – an OEC – that is a weighted combination of such objectives (Roy 2001, 50, 405–429) may be the more desired solution. And like metrics overall, ensuring that the metrics and the combination are not gameable is critical (see *Sidebar: Gameability* in Chapter 6). For example, basketball scoreboards don't keep track of shots beyond the two- and three-point lines, only the combined score for each team, which is the OEC. FICO credit scores combine multiple metrics into a single score ranging from 300 to 850. The ability to have a single summary score is typical in sports and critical for business. A single metric makes the exact definition of success clear and has a similar value to agreeing on metrics in the first place: it aligns people in an organization about the tradeoffs. Moreover, by having the discussion and making the tradeoffs explicit, there is more consistency in decision making and people can better understand the limitations of the combination to determine when the OEC itself needs to evolve. This approach empowers teams to make decisions without having to escalate to management and provides an opportunity for automated searches (parameter sweeps).

If you have multiple metrics, one possibility proposed by Roy (2001) is to normalize each metric to a predefined range, say 0–1, and assign each a weight. Your OEC is the weighted sum of the normalized metrics.

Coming up with a single weighted combination may be hard initially, but you can start with classifying your decisions into four groups:

1. If all key metrics are flat (not statistically significant) or positive (statistically significant), with at least one metrics positive, then ship the change.

2. If all key metrics are flat or negative, with at least one metric negative, then don't ship the change.
3. If all key metrics are flat, then don't ship the change and consider either increasing the experiment power, failing fast, or pivoting.
4. If some key metrics are positive and some key metrics are negative, then decide based on the tradeoffs. When you have accumulated enough of these decisions, you may be able to assign weights.

If you are unable to combine your key metrics into a single OEC, try to minimize the number of key metrics. Pfeffer and Sutton (1999) warn about the Otis Redding problem, named after the famous song "Sitting by the Dock of the Bay," which has this line: "Can't do what ten people tell me to do, so I guess I'll remain the same." Having too many metrics may cause cognitive overload and complexity, potentially leading the organization to ignore the key metrics. Reducing the number of metrics also helps with the multiple comparison problems in Statistics.

One rough rule of thumb is to try to limit your key metrics to five. While using a strong 0.05 p-value threshold by itself can be abused – p-hacked, if you will (Wikipedia contributors, Multiple Comparisons problem 2019)– we can still use the underlying statistical concept as a way to understand this heuristic. Specifically, if the Null hypothesis is true (no change), then the probability of a p-value < 0.05 for a single metric is 5%. When you have k (independent) metrics, the probability of having at least one p-value < 0.05 is $1-(1-0.05)^k$. For $k=5$, you have a 23% probability of seeing something statistically significant. For $k=10$, that probability rises to 40%. The more metrics you have, the higher the chance that one would be significant, causing potential conflicts or questions.

One final benefit of an OEC that is agreed upon: you can automatically ship changes (both simple experiments and parameter sweeps).

Example: OEC for E-mail at Amazon

At Amazon, a system was built to send e-mails based on programmatic campaigns that targeted customers based on various conditions, such as (Kohavi and Longbotham 2010):

- Previously bought books by an author with a new release: A campaign e-mailed them about the new release.
- Purchase history: A program using Amazon's recommendation algorithm sent an e-mail like this: "Amazon.com has new recommendations for you based on items you purchased or told us you own."

- Cross-pollination: Many programs were very specific and defined by humans to e-mail product recommendations to customers who bought items from specific combinations of product categories.

The question is what OEC should be used for these programs? The initial OEC, or "fitness function," as it was called at Amazon, gave credit to a program based on the revenue it generated from users clicking-through on the e-mail.

The problem is that this metric is monotonically increasing with e-mail volume: more campaigns and more e-mails can only increase revenue, which led to spamming users. Note that this property of increasing revenue with e-mail volume is true even when comparing revenue from the Treatment users (those receiving the e-mail) to Control users (those who don't).

Red flags went up when users began complaining about receiving too many e-mails. Amazon's initial solution was to add a constraint: a user can only receive an e-mail every X days. They built an e-mail traffic cop, but the problem was that it became an optimization program: which e-mail should be sent every X days when multiple e-mail programs want to target the user? How could they determine which users might be open to receiving more e-mails if they found them truly useful?

Their key insight was that the click-through revenue OEC is optimizing for short-term revenue instead of user lifetime value. Annoyed users unsubscribe and Amazon then loses the opportunity to target them in the future. They built a simple model to construct a lower bound on the user lifetime opportunity loss when a user unsubscribes. Their OEC was:

$$\text{OEC} = \left(\sum_i Rev_i - s * unsubscribe\_lifetime\_loss \right) \Big/ n$$

where:

- i ranges over e-mail recipients for the variant
- s is the number of unsubscribes in the variant
- unsubscribe_lifetime_loss is the estimated revenue loss of not being able to e-mail a person for "life"
- n is the number of users in the variant.

When they implemented this OEC with just a few dollars assigned to unsubscribe lifetime loss, more than half of the programmatic campaigns were showing a negative OEC!

More interestingly, the realization that unsubscribes have such a big loss led to a different unsubscribe page, where the default was to unsubscribe from this

"campaign family," not from all Amazon e-mails, drastically diminishing the cost of an unsubscribe.

Example: OEC for Bing's Search Engine

Bing uses two key organizational metrics to measure progress: query share and revenue, as described in *Trustworthy online controlled experiments: Five puzzling outcomes explained* (Kohavi et al. 2012). The example shows how short-term and long-term objectives can diverge diametrically. This problem is also included in *Data Science Interviews Exposed* (Huang et al. 2015).

When Bing had a ranker bug that resulted in very poor results being shown to users in a Treatment, two key organizational metrics improved significantly: distinct queries per user went up over 10%, and revenue-per-user went up over 30%. What should the OEC for a search engine be? Clearly, the search engine's long-term goals do not align with these two key metrics in experiments. If they did, search engines would intentionally degrade quality to raise query share and revenue!

The degraded algorithmic results (the main search engine results shown to users, also known as the 10 blue links) forced people to issue more queries (increasing queries-per-user) and click more on ads (increasing revenue). To understand the problem, let's decompose query share:

Monthly *query share* is defined as distinct queries for the search engine divided by distinct queries for all search engines over one month. Distinct queries per month decomposes to the product of these three terms as shown in Equation 7.1:

$$n \frac{Users}{Month} \times \frac{Sessions}{User} \times \frac{Distinct\ queries}{Session}, \tag{7.1}$$

where the second and third terms in the product are computed over the month, and a session is defined as user activity that begins with a query and ends with 30 minutes of inactivity on the search engine.

If the goal of a search engine is to allow users to find their answer or complete their task quickly, then reducing the distinct queries per task is a clear goal, which conflicts with the business objective of increasing query share. As this metric correlates highly with distinct queries per session (more easily measurable than tasks), distinct queries alone should not be used as an OEC for search experiments.

Given the decomposition of distinct queries shown in Equation 7.1, let's look at the three terms:

1. Users per month. In a controlled experiment, the number of unique users is determined by the design. For example, in an A/B test with 50/50 split, the number of users that fall in each variant will be approximately the same, so you cannot use this term as part of the OEC for controlled experiments.
2. Distinct queries per task should be minimized, but it is hard to measure. You can use the metric distinct queries per session as a surrogate; however, this is a subtle metric because increasing it may indicate that users have to issue more queries to complete the task but decreasing it may indicate abandonment. Thus, you can aim to decrease this metric as long as you also check that the task is successfully completed (i.e., abandonment does not increase).
3. Sessions-per-user is the key metric to optimize (increase) in controlled experiments. Satisfied users visit more often.

Revenue per user should likewise not be used as an OEC for search and ad experiments without adding other constraints. When looking at revenue metrics, we want to increase them without negatively impacting engagements metrics. A common constraint is to restrict the average number of pixels that ads can use over multiple queries. Increasing revenue per search given this constraint is a constraint optimization problem.

Goodhart's Law, Campbell's Law, and the Lucas Critique

The OEC must be measurable in the short term (the duration of an experiment) yet believed to causally drive long-term strategic objectives. Goodhart's law, Campbell's law, and the Lucas Critique all highlight that correlation does not imply causation and that in many situations organizations that pick an OEC are fooled by correlations.

Charles Goodhart, a British economist, originally wrote the law: "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (Goodhart 1975, Chrystal and Mizen 2001). Today it's more common to reference Goodhart's law as: "When a measure becomes a target, it ceases to be a good measure" (Goodhart's law 2018, Strathern 1997).

Campbell's law, named after Donald Campbell, states that "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell's law 2018, Campbell 1979).

Lucas critique (Lucas critique 2018, Lucas 1976) observes that relationships observed in historical data cannot be considered structural, or causal. Policy decisions can alter the structure of economic models and the correlations that held historically will no longer hold. The Phillips Curve, for example, showed a historical negative correlation between inflation and unemployment; over the study period of 1861−1957 in the United Kingdom: when inflation was high, unemployment was low and vice versa (Phillips 1958). Raising inflation in the hope that it would lower unemployment assumes an incorrect causal relationship. As a point in fact, in the 1973–1975 US recession, both inflation and unemployment increased. In the long run, the current belief is that the rate of inflation has no causal effect on unemployment (Hoover 2008).

Tim Harford addresses the fallacy of using historical data by using the following example (Harford 2014, 147): "Fort Knox has never been robbed, so we can save money by sacking the guards." You can't look just at the empirical data; you need also to think about incentives. Obviously, such a change in policy would cause robbers to re-evaluate their probability of success.

Finding correlations in historical data does not imply that you can pick a point on a correlational curve by modifying one of the variables and expecting the other to change. For that to happen, the relationship must be causal, which makes picking metrics for the OEC a challenge.