**A/B testing**

<https://exp-platform.com/advanced-topics-in-online-experiments/>

**The Overall Evaluation Criterion (OEC)**

<https://onedrive.live.com/view.aspx?resid=8612090E610871E4!282179&ithint=file%2cdocx&app=Word&authkey=!ANFGOBrhVt91ODk>

When designing controlled experiments, one of the most important questions to ask early is: what are you optimizing for?

1. Agreeing on a goal and metric to measure is critical.
2. Sensitivity depends on three factors: the statistical variance of the underlying metric, the effect size (delta between treatment and control in an experiment), and the number of experiment units (e.g., users). A whole-page click-through metric, a measure of “success” (e.g., purchase), and time to success, are usually good key metrics that are sensitive enough for experimentation.
3. Improvements to the key metrics should have a causal relationship to the key organizational long-term goals. One suggestion that we have is to think of customer lifetime value as a guiding principle for key metrics.
4. Expect to define your key metrics and modify them over time, as experiments shed light on the strengths and weaknesses of metrics.
5. Sessions/user is the key metric to optimize (increase) in experiments, as satisfied users will come more. And it could be used as a OEC(Overall Evaluation Criteria) for a search engine.
6. Look for other constraints when trying to define the OEC for a search engine. For example: we would love to increase the revenue per user through ad display. At the same time, we do not want to negatively impact the user engagement metrics.
7. correlation does not imply causation. 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 is why picking metrics for the OEC is so hard.
8. In practice, improvements to key metrics are achieved by many small changes: 0.1% to 2%. <https://onedrive.live.com/view.aspx?resid=8612090E610871E4!323967&ithint=file%2cdocx&app=Word&authkey=!APyJuF_t0dOFj_M>

**A/A Test** <https://onedrive.live.com/view.aspx?resid=8612090E610871E4!288827&ithint=file%2cdocx&app=Word&authkey=!AE3UclwDsmPl80Y>

Running an A/A Test, or rather many of them, is a critical part of establishing trust in an experimentation platform for a given domain. The test has been so useful because it fails so many times in practice, leading to a re-evaluation of assumptions, and helping to identify software bugs. The idea is simple: take an A/B testing system, split the users into two groups, but make B identical to A (hence the name A/A test). If the system is operating correctly, then in repeated trials about 5% of the time a given metric will be statistically significant with p-value less than 0.05. More generally, for each non-discrete metric the distribution of p-values from repeated trials should be close to a uniform distribution.

**Benefits of running A/A test**

1. The test can be used to check whether splits are happening according to the planned percentages. If you are running a 20%/20% experiment, are you getting approximately equal percentages? A sample-ratio mismatch test can be used.

2. Is the data collected matching the system of record? If the A/B testing system uses a different logging system, then are key metrics matching?

a. If the system of records shows X users visited the website during the duration of the experiment, and you ran Control and Treatment at 20% each, do you see about 20% of X in each? Are you leaking users someplace? This can create a bias.

b. If the system of record shows R revenue was generated, does the testing system show this? If you ran a 50/50% experiment, your revenue should exactly match, as each user is either in Control or Treatment.

3. The A/A test can provide you with variances of metrics so you can do Statistical Power calculations to determine how long to run your A/B tests for a given minimal detectable effect.

**Click Through Rate**

Click-through rate is a very common metric, but there are two common ways to compute it.

Here are two reasonable definitions of clickthrough rate.

1. If we have two users, one with no clicks and a single page view, and the other with 2 click, one on each of the two page views, then 𝐶𝑇𝑅1= (0+2)/1+2= 23 i.e., summation of clicks per user divided by summation of page views per user.
2. The second method is to first average each user’s clickthrough rate, and then average all the clickthrough rates, essentially a double average as follows:

Using the example above, we have 𝐶𝑇𝑅2= ((0/1)+(2/2))/2= 1/2

There is no right or wrong above: they are two useful definitions of CTR that yield different results due to different ways of averaging users. In our scorecards, we expose both metrics, although we generally find the latter to be more robust to outliers and prefer it.

The computation of the variance is where mistakes are made. If the A/B test is randomized by user, as is commonly done, then computing the variance of the first definition as

VAR(𝐶𝑇𝑅1)= ∑∑((𝑋𝑖𝑗−𝐶𝑇𝑅1)^2)/N^2

is biased, as it assumes that the 𝑋𝑖𝑗’s are independent, which is inaccurate.

**How long should you run an A/B test on your site before you declare one a winner?**

One point to note is that the key metric (Overall Evaluation Criteria) is really important and the number of users to be allocated to each variant differs depending on whether the metric is a low varying or high varying one. Therefore optimizing for the metric of interest matters a lot.

When the results/outcome could be misleading:

* The OEC, or Overall Evaluation Criterion, is not properly taken into account. It’s OK to run small experiments for ideation, but they can never ship without first being powered up sufficiently.
* Don’t stop early because something is statistically significant.
* The power formula assumes a stationary metric value. Plot the metric (e.g., conversions) and you’ll notice that it’s not stationary.  
  It usually varies throughout the day. It may be different over the weekend?
* Are you interested in segmenting your data afterwards? Looking at desktop vs. mobile? Perhaps segmenting by browsers? You’ll need more users, which means running the experiment longer or (better) at larger percentages.
* Bots and outliers must be removed.
* Novelty (newness) and primacy effects.
* In online experiments, not all users join at the same time. For example, Sessions/User is monotonically increasing for each user, but the mix of users that joined early and users admitted later complicates the distribution, and running the experiment longer does not help.
* Ramp-up time and Simpson’s paradox.
* Is the metric homoscedastic (same variance over time)? Plot the metric over time to check this

**Point to be noted:**

* True effect size and delta (delta=mu1-mu2 where mu1 could be control’s and mu2 could be treatment’s) represent one and the same.
* When getting the true metric is hard, sites will use a surrogate metric

**Flow of an A/B test**

1. Set Null hypothesis
2. Choose your Confidence level
3. Do Power Analysis
4. A/A test
5. Standard deviation
6. Standard error
7. Statistical significance

**Overall Evaluation Criteria and statistical significance**

Statistical Significance. To evaluate whether the Overall Evaluation Criterion differs for user groups exposed to Treatment and Control variants, a statistical test can be done. If the test rejects the null hypothesis, which is that the OECs are not different, then we accept a Treatment as being statistically significantly different.

For example: Say, we have an existing design as control and a new design as treatment. We have decided to run a controlled experiment to check whether the new design would bring in more clicks on the revenue generating click buttons. Turns out the new design ended up with fewer clicks and thus a lower click through rate compared to that of the control. But it could be devastating to ignore the fact the new design brought in more qualifying customers to actually buy than the control. Also, that the revenue actually generated via the new design is higher than that via the control. So just focusing on the local change constrained to a small area might not have the big picture view re the OEC like the whole page click through rate.

Online surveys are not a suitable source of input for Overall Evaluation Criteria.

Setting the OEC to time spent on page (dwell time) fails the litmus test noted in pitfall 1. For example, in a Microsoft health related site, a widget was redesigned to make health articles more accessible. Time spent on pages and total session time increased (satisfying the objective), but drilling down to the reasons, the new widget in the Treatment was used less often than the one in the Control. Users may have been more confused, thus taking longer to find what they need.

For example, a particularly successful support site experiment we ran involved the test of a rudimentary personalization feature. The support.microsoft.com site contained a top center “Instant Answers” module with links to common support issues selected by the site editors. We tested a new treatment that personalized these links by the browser and operating system versions of the user’s HTTP header. The treatment performed over 50% better than the control on the OEC of Click-through rate, without decreasing the clickthrough rate for the whole page.

**Confidence intervals**

For many online metrics, the difference in the means is so small that percent change has much more intuitive meaning than the absolute difference. For example, for a recent experiment we ran, the Treatment effect for clickthrough rate was 0.00014. This translated to a 12.85% increase for the Treatment. The latter number was much more meaningful to decision makers. The percent difference is calculated as the delta between the means of the Treatment and Control divided by the mean for the Control times 100%.

These formulas assume the covariance between the Treatment and Control mean is zero, which will be true in a controlled experiment when the randomization is carried out properly.

OECs may be a combination of metrics, or key performance indicators (KPIs). This combination could be either

1) A linear combination of metrics

2) A nonlinear combination of metrics that have the same basis (Two metrics have the same basis if they are calculated over the same experimental unit. For example, page views per user-day and click throughs per user-day have the same basis, user-day.)

3) A nonlinear combination of metrics that do not have the same basis.

In the first case, the mean and variance of the OEC can be calculated from the means and variance of the metrics using the standard formulas and the confidence intervals are the usual symmetric confidence intervals using the normal distribution.

In the second case, one can calculate the OEC for each experimental unit then calculate the mean and variance of the OEC values across experimental units and then the confidence intervals.

The third case is more challenging. If the OEC is a general function of k primary metrics, i.e. OEC = g(X1, X2, ..., Xk), and if g(.) is a totally differentiable function of k variables, if (X1, X2, ..., Xk) asymptotically follow a joint Normal distribution with means μ1, μ2,... μk, and covariances σij, i, j = 1,...k, then the OEC will asymptotically follow a Normal distribution with mean g(μ1, μ2,... μk) and variance. We also have to assume the sample sizes are large enough for g(X1, X2, ..., Xk) to have a Normal distribution. (check the link for the possible formula: <https://ai.stanford.edu/~ronnyk/2009-ExPpitfalls.pdf>)

**Pitfall 2:** Incorrectly computing confidence intervals for percent change and for OECs that involve a nonlinear combination of metrics

**Pitfall 1:** Picking an OEC for which it is easy to beat the control by doing something clearly “wrong” from a business perspective.

**METRICS, STANDARD DEVIATIONS AND POWER**

We discovered that variances for some metric families are inaccurately estimated using the standard statistical formulas. Specifically, the variance for click-through rate (CTR), defined as (sum of clicks)/(sum of page views) for the Treatment or Control for the time period of the experiment was significantly underestimated. In these cases, we have found the Bootstrap method (Efron, 1993) to be an excellent way to estimate the variance.

We currently take 1000 bootstrap samples. We recommend that you compare the formula variance for any metric with the Bootstrap estimate if you are not sure the formula for the variance is accurate. The metrics may be considered in two categories: those where the experimental unit is the same as the randomization unit and those where it is not. (for more details refer: <https://ai.stanford.edu/~ronnyk/2009-ExPpitfalls.pdf>)

There is usually some positive correlation between experimental units for metrics like: page views per user per day, page views per session, etc., and sites that have more loyal customers (higher return rate) have higher correlations. Ignoring the correlations leads to underestimation of the standard deviation. We have been using Bootstrapping to estimate the standard deviation for these metrics and getting good results, validated through A/A tests.

The only class of metrics where the power and standard deviation calculations are straightforward are conversion rates for users. For example, the percent of users who purchase an item or the percent of users who click on a link. These metrics follow the Bernoulli distribution when randomization is by user

**Pitfall 3:** Using standard statistical formulas for computations of variance and power.

**SIMPSON’S PARADOX**

Problems due to:

* Ramp-up
* Different proportions for different periods

While occurrences of Simpson’s paradox are unintuitive, they are not uncommon, and we have seen them happen multiple times in real life. Possible solutions include: (i) paired t-tests where each pair (Control, Treatment) is chosen from a period where the proportions were stable; and (ii) using weighted combinations. The simplest solution, which we use, is to throw away the data from the ramp-up period, which is usually short relative to the experiment.

**Pitfall 4:** Combining metrics over periods where the proportions assigned to Control and Treatment vary, or over subpopulations sampled at different rates

**ROBOTS IMPACT RESULTS**

For experimentation, we are primarily concerned with removing robots that cause a bias. If the traffic from a robot is distributed across the variants of an experiment in an unbiased way, then the presence of the robot adds noise to the data and reduces the power of the experiment but does not invalidate the results. Robots that are seen as multiple unique users due to resetting their cookies or running from multiple machines do not introduce bias. Robots that act like a single user and consistently generate traffic for a single variant, however, can create a significant bias. For example, if a robot consistently assigned to variant A generates an excessive number of clicks, it may cause A to have a statistically significantly higher click-through rate than B even if B is preferred by human users.

The novel evaluation scheme we propose is to use A/A tests, where users are split into Control and Treatment, but there is no systematic difference between the two versions they are exposed to. The Null hypothesis in an A/A test should be rejected about 5% of the time when a 95% confidence level is used. If this does not hold true, then there is a bias introduced by extreme behavior of users, which are most likely robots being assigned to a particular variant. Multiple A/A tests must be run in order to have confidence whether biased robots exist in the data. However, an interesting observation is that these don’t have to be live A/A tests. It is sufficient to run tests post-hoc ("offline") by re-randomizing users and assigning them to Control/Treatment and evaluating the hypothesis that they are the same.

**Pitfall 5:** Neglecting to filter robots

**AUDITING THE ANALYSES**

1. Logging Test

Compare with system of record

Compare with generated data

Look for unexpected patterns: Volume of data over time, Number of new and repeat users over time, Ratios ofrelated observations over time, and Dimensional analysis

1. A/A Test
2. Offline A/A Test
3. Rich Instrumentation: Collect data at referrer and destination points, Over-instrumenting is better than under- instrumenting

**A/A Test**

The application code used to assign users to variants and execute the appropriate variant must be the same as it would if the variants were different. Running an experiment in this configuration allows us to perform a number of sanity checks to validate that the experimentation apparatus itself is functioning properly.

Verifying that each end user consistently received a single variant can be done by injecting variant specific information into the user behavior data. For example, if users in variant A should receive page X but users in variant B should receive page Y then recording the URL (X or Y) in a page view observation allows checking whether any user received the wrong page.

Recommended to use 50% of users for each variant of an experiment to maximize the statistical power in A/B tests.

**Offline A/A Test**

Offline A/A test as a mechanism to evaluate robot detection algorithms. It can also be useful in uncovering other metric calculation problems as well.

It is important to note that offline A/A tests identify very different problems than normal A/A tests. An offline A/A test finds problems with the calculation of metrics whereas a normal A/A test detects variant assignment bugs and biased data collection.

**Rich Instrumentation:** Collect data at referrer and destination points

Over-instrumenting is better than under- instrumenting: For example, by collecting server side page request observations we were able to identify an issue in which FireFox was requesting each page twice due to an IMG tag with an empty SRC attribute on the page.

**Pitfall 6:** Failing to validate each step of the analysis pipeline and the OEC components

**Control is Crucial**

Our recommendation is to choose an approach to experimentation that does not require a redirect, but if you need to use that method you should include the redirect in all variants you are testing.

**Pitfall 7:** Forgetting to control for all differences, and assuming that humans can keep the variants in sync

**What is important is to review the factors that impact the test:**

1. Confidence level. Commonly set to 95%, this level implies that 5% of the time we will incorrectly conclude that there is a difference when there is none (Type I error). All else being equal, increasing this level reduces our power (below).

2. Power.Commonlydesiredtobearound80–95%, although not directly controlled. If the Null Hypothesis is false, i.e., there is a difference in the OECs, the power is the probability of determining that the difference is statistically significant. (A Type II error is one where we retain the Null Hypothesis when it is false.)

3. Standard error. The smaller the Std-Err, the more powerful the test. There are three useful ways to reduce the Std-Err:

a. TheestimatedOECistypicallyameanoflargesamples.AsshowninSect.3.1, the Std-Err of a mean is inversely proportional to the square root of the sample size, so increasing the sample size, which usually implies running the exper- iment longer, reduces the Std-Err and hence increases the power for most metrics.

b. Use OEC components that have inherently lower variability, i.e., the Std- Dev, σ , is smaller. For example, conversion probability (0–100%) typically has lower Std-Dev than number of purchase units (typically small integers), which in turn has a lower Std-Dev than revenue (real-valued).

c. Lower the variability of the OEC by filtering out users who were not exposed to the variants yet were still included in the OEC. For example, if you make a change to the checkout page, analyze only users who got to the page, as everyone else adds noise, increasing the variability.

*4. Effect*. The difference in OECs for the variants, i.e. the mean of the Treatment minus the mean of the Control. Larger differences are easier to detect, so great ideas will unlikely be missed. Conversely, Type II errors are more likely when the effects are small.

**Define power:**

Commonly desired to be around 80–95%, although not directly controlled. If the Null Hypothesis is false, i.e., there is a difference in the OECs, the power is the probability of determining that the difference is statistically significant. (A Type II error is one where we retain the Null Hypothesis when it is false.)