Use minimum detectable effect to prioritize experiments

Skip Ahead

- How MDE affects sample size
- Set boundaries by estimating MDE
- MDE and operation constraints

This article will help you:
- Estimate the time it will take to run an experiment, based on the lift you expect to measure
- Prioritize an experiment in your roadmap based on ROI

Many optimization programs prioritize their roadmaps by estimating the effort versus impact of an individual experiment. But estimating ROI before you run an experiment can feel like guesswork. Use the minimum detectable effect (or MDE) to gauge the potential effort and impact of an experiment.

The minimum detectable effect represents the relative minimum improvement over the baseline that you're willing to detect in an experiment, to a certain degree of statistical significance. It can help you figure out the likely relationship between impact and effort - or cost and potential value - for your experiment.

Use it to benchmark how long to run an experiment and the impact you're likely to see, so you can prioritize experiments according to expected ROI. Set expectations for how long it may take to run an experiment based on MDE, depending on how granular you'd like your results to be.

MDE can help provide a guideline through the uncertainty of experimentation, so you can prioritize your roadmap effectively. Read on to learn about MDE for prioritization.

How MDE affects sample size

One major cost in every experiment is the time it takes to reach a statistically significant result. To estimate how long a given experiment will need to run to achieve statistical significance, you need to determine the following:
• **Traffic allocated for the experiment**
  What percentage of your traffic (or unique weekly visitors) will you allocate for this experiment?

• **Total sample size**
  Total sample size is the number of variations multiplied by the sample size per variation. Use Optimizely’s sample size calculator to estimate the sample size you’ll need to reach statistical significance, depending on the baseline conversion rate and the MDE.

Use the following formula to determine how long to run an experiment:

\[
\text{Total sample size} \times \text{Number of weeks} = \text{Sample size needed} = \text{Traffic for test per week} = \# \text{of weeks to run test}
\]

Once you divide **total sample size** by **traffic allocated to the experiment, per week**, you’ll know approximately how long it takes to run this experiment -- and prioritize accordingly. For most organizations, the most difficult part of this calculation is MDE.

To calculate the **sample size per variation** for your experiment, you need the current baseline conversion rate and the **minimum detectable effect**. The MDE represents the relative minimal improvement over the baseline that you’re willing to detect in this experiment to a certain significance level; for now, let’s just assume a standard 95% statistical significance.

If your experiment measures an actual improvement that is **equal to or higher than** the MDE, you’ll reach significance within given sample size. In other words, you’ll see a significant result with equal or fewer visitors than originally estimated -- and you can call a winner more quickly. However, if your experiment detects improvement at a level that’s lower than the MDE you set, it won’t reach statistical significance within the given sample size. You’d have to keep running the experiment to call a winner.

Imagine, for example, that you’re running an experiment to optimize a checkout flow. You measure conversions with a pageview goal on the checkout confirmation page; the baseline conversion rate is 10%. You estimate that your variation will improve the baseline by at least 5% and your variation conversion rate will be 10.5% or greater -- so, your MDE is 5%.
With the help of Optimizely’s sample size calculator, you determine a sample size of 62,000 per variation. In this experiment, which includes one original and one variation, you’d need approximately 124,000 visitors to detect a change of 5% or more at 95% statistical significance. So, you launch the experiment.

Once results start flowing in, you note that the actual conversion rate for your variation is higher than 10.5%. If this trend continues, it’s likely that your experiment will reach significance within 124,000 visitors. But if the conversion rate is lower than 10.5% - meaning the improvement is less than the 5% originally predicted - you probably won’t reach statistical significance by 124,000 visitors.

At this point, you’d decide whether to keep running the experiment. Depending on your results, you may decide to gather more data or move on to the next idea.

**Set boundaries by estimating MDE**

Rather than trying to get your MDE exactly right, use it to set boundaries for your experiment so you can make informed business decisions. With a more nuanced understanding of how MDE affects sample size and goals, you can decide when to keep running an experiment, given certain operational constraints.

Notice how the baseline conversion rate and MDE directly affect the sample size:

The smaller your baseline is, the larger the sample size required to detect the same relative change (MDE).

<table>
<thead>
<tr>
<th>Baseline</th>
<th>MDE</th>
<th>Statistical significance</th>
<th>Sample size (per variation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15%</td>
<td>10%</td>
<td>95%</td>
<td>7,271</td>
</tr>
<tr>
<td>10%</td>
<td>10%</td>
<td>95%</td>
<td>12,243</td>
</tr>
<tr>
<td>3%</td>
<td>10%</td>
<td>95%</td>
<td>51,141</td>
</tr>
</tbody>
</table>

The smaller your MDE is, the larger the sample size required to reach statistical significance.
<table>
<thead>
<tr>
<th>Baseline</th>
<th>MDE</th>
<th>Statistical significance</th>
<th>Sample size (per variation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>10%</td>
<td>95%</td>
<td>12,243</td>
</tr>
<tr>
<td>10%</td>
<td>5%</td>
<td>95%</td>
<td>59,401</td>
</tr>
<tr>
<td>10%</td>
<td>3%</td>
<td>95%</td>
<td>185,661</td>
</tr>
</tbody>
</table>

The graph below clearly demonstrates how your sample size may balloon as you attempt to detect a smaller MDE.

Sample size translates directly into **how long it takes to run an experiment**.

number of weeks to run an experiment = total sample size / unique visitors per week

Let's return to the example above, where the **baseline** was 10% and **MDE** was 5%. If you had 40,000 weekly unique visitors on the page you’re planning to experiment, you'd calculate the following:

total sample size / unique visitors per week = 60,000 / 40,000 = 1.5, or approximately **2 weeks**

**Note:**

Since most business metrics run on a weekly cycle, consider rounding up the time you calculate for running an experiment to **whole weeks**.

The baseline, number of variations, number of unique visitors, and statistical significance are constant for this experiment. So, you can plot **the time it takes to run this experiment as a function of the MDE**.

Notice that an attempt to detect improvement to a granularity of 4% lift or less will take **at least five weeks**.

Now, you can use this information to prioritize this experiment. If the 4% lift on this goal moves the needle on an important metric and the estimated time to run this experiment is within a realistic range for your roadmap, you may want
to move forward with the experiment. For example, 4% lift in 5 weeks may be a reasonable tradeoff of impact (lift) for effort (experiment runtime) -- but 2% lift measured over several months may not be. If the traffic cost is too high, consider de-prioritizing the hypothesis in your **roadmap** or seeking to measure lift with less granularity.

Use a range of MDEs to get a feel for the time you’re willing to invest into each experiment. This range can also enable you to decide to keep running or stop an experiment with inconclusive results, when you evaluate results.

**MDE and operation constraints**

Sometimes, your time to run an experiment may be **limited for operational reasons**, such as:

- **Traffic** - you can only allocate a limited amount of traffic to an experiment (or, your site has relatively low traffic)
- **Time** - you’re trying to get results quickly due to operational pressures
- **Value** - you won’t run the experiment unless you can prove that it provides a certain amount of value

Use MDE to make informed business decisions.

If you have only two weeks to run an experiment, for example, plot how MDE can be measured for other goals in the experiment. See what impact you can observe in the two weeks that you have.

According the sample graph of Weeks to run versus MDE above, you’d be able to capture very small changes for **Goal 2** (yellow) and **Goal 3** (green) within two weeks: 2% and 3% lift respectively. However, you’d only be able to observe an 8% change or higher in **Goal 4** (red).

**Is 8% lift granular enough for this experiment?** If so, launch the experiment with the two-week window. If you need to detect improvement in that goal with better granularity, consider pushing for more time for the experiment or blocking out time for it in your roadmap. In either case, you have a clearer idea of what to expect from this experiment.

Use MDE and time to run an experiment to inform how you prioritize experiments in your roadmap.

Eventually, getting better at estimating MDEs is a matter of setting limits and ranges rather than looking for exact numbers. Often, if you are off by a few percent, you can run the experiment for a while longer to find the answer.

But always ask yourself:

- What other experiments could I be running instead?
• Am I devoting my resources to the right experiment?

Examine your experiment plan to understand how the estimated time that an experiment will take can provide insight into how it should be prioritized. An MDE-based approach to prioritization and planning can help you build a more detailed, intentional roadmap to experimentation.