How To Write A Test Spec

Evan Bloom

A Test Spec is a document often used to organize a test. It is a living document, meant to be edited as you and your team plan an experiment. You will inevitably fill in more details as you go. This document follows the outline of a test spec and provides commentary or examples of how to fill it in.

Often the Context, Hypotheses, and Decision Framework sections are meant to be communicated outwards. People on other teams, not close to your work, should be able to get a broad sense of what the test is addressing and how it will work. In this way, you can both make others aware of what's happening (to understand how it may interact with other experiments or plans within your organization) and elicit feedback from others.

The Scope, Cell design, and Metrics sections will often be communicated inwards to scientists, engineers, product managers, and others involved in the execution of the test. This is critically important, to organize what happens and ensure the intervention is actually what you intended.

Name of Your Test

(something descriptive, bonus points if witty or includes a pun)

Summary

This is a document that describes what a *Test Spec* might look like. There should be a summary highlighting the most important points.

- The main hypothesis should go here
- You should also state the decision that the test is designed to inform
- You might state when you expect to run the experiment and make a decision

Details

Context

This is where you describe the opportunity you are trying to address.

- Explain the current situation for the business or organization

- Describe what might change as a result of this test
- Estimate why this is worth doing. Use some data if you have it, even though it won't be perfect.
    - How many people will it affect?
    - How big an outcome will it drive?
    - Why is this material for your organization?
- Help the reader understand the broader landscape
    - Are there other tests, experiments, projects related to this experiment?
    - Is this similar to things that have been tried before, inside or outside your organization? What have you learned?
    - Is there literature on this?

Hypotheses

In the context section, you probably spoke generally and broadly. Now it’s time to be specific. An experiment can have multiple hypotheses. They should be called out separately and uniquely.

- Often, there is one cell of your test per hypothesis.
- Sometimes you might have multiple cells examining the same hypothesis (for example, the intensity or dosage of the treatment).
- Or you might need multiple cells to measure a single hypothesis (for example, you want to measure a relationship to dosage).
- Sometimes by comparing different combinations of cells to each other, you can learn about many hypotheses (technically 4 cells creates 4! comparisons, but many won't be very relevant).
- There is no shame in keeping it simple. Many hypotheses can be a sign that you are trying to do too much at once.

Here is what a hypothesis might look like:

- What variable are you changing? What do you expect to happen because of it?
    - It may be good to be clear what the comparison is against.
- Often it is good to have a second point to the hypothesis that explains the mechanism behind the test.
- For example: *If we send more emails to users they will visit our site more often*
    - *More emails* is what we are changing and *visit our site more often* is the outcome we expect
    - We might follow up with: *We know this is working because users clicked links in the emails or We will expect to see users open our emails more frequently.*
        - These sub-hypotheses are good, because it would mean that we might be skeptical of our test result if we saw an increase in site visits, but did not see users opening additional emails.
    - You should always investigate the specificity of your hypothesis. In this case, what does “more” really mean? Are you sending more emails to the same people (increasing the intensive margin) or a new audience (extensive margin).

        - You can also do better by defining what are you comparing against: no emails or the current email experience?
            - If the latter, is that well defined? Do you have a system or policy in place, or does some person make individual decisions about what they are sending)

Usually we just list them

- **H1:** this is a hypothesis
- **H2:** this is another hypothesis

Decision Framework

Tests are generally designed to inform decisions. So every test spec should be as specific as possible to describe what decision we will inform:

1. Whether we will **roll-out the user experience** that we tested to all users
2. Other times, we might **learn a parameter** that feeds into another model (or decision process) that then controls the outcome
    a. *For example:* knowing the rate at which there are diminishing returns to spend on ads might inform a total budget allocation.
3. It may be a set of follow up tests and investments we will make to improve further. This might help us gauge the **size of an opportunity**, prior to making the full investment
    a. *For example:* we might want to learn if Facebook Ads can get people to vote by mail. If effective, then we might want to begin a program of learning who to target and what the content of the ads should be.
    b. In these cases, it might be better to design your test considering an upper or lower bound that would improve with further iteration.
4. Other times, the exact experience doesn't matter very much, it was designed to manipulate something which we then have other levers to control by **simulating a strategy**.
    a. *For example:* Spotify might promote only a subset of Podcasts to a set of users. This would be meant to approximate signing more or fewer podcasters to exclusive contracts.
        i. Finding the right way to increase or decrease the number of podcasts Spotify promotes might be different than if they were simply trying to optimize their home page for the number of podcasts vs volume of music.

Based on this, we can think about how the data coming out the test will inform that decision:

1. Do you need to see a **statistically significant win** to make a change?
    a. We might do this when we believe our status-quo is already pretty strong. We would then think we need good evidence to change it
    b. If you can be very precise, is there some higher bar than statistically significance? Sometimes small significant changes are not **materially significant,** particularly as there often implicit or explicit costs to changing. In such cases it would be good to set a bar for a materially significant effect size.

2. Do we need to not see statistically significant loss, or **do-no-harm**?
    a. We might want to do this when upgrading to something that has other benefits (cheaper to maintain, easier to build upon) and we want to rollout so long as it maintains our metic.
3. Are there **multiple metrics** that we care about?
    a. For example: benefit vs cost?
    b. If so, How do we weigh them?
        i. Is there a way to combine them?
        ii. Are we looking for a pareto-improvement (improve one metric without harming the other)?
4. Are there **risks outside of the test** we also want to monitor?
    a. For example, is this difficult to execute on, and causes others in your organization and increased workload?
    b. It's good to be explicit about other things, besides your primary test hypothesis that you will monitor, and weigh into your final decision.

Nearly every decision is about weighing tradeoffs. It can be overwhelming to list literally every scenario (especially if there are multiple metrics and multiple other things you will monitor), but it can be good to summarize the high level framework of making a decision.

A simple table like this, that weights some tradeoffs can be helpful. A decision tree might work well too. Play with different ways to summarize your thinking, in a way that is straight-forward for people outside your test.

Metric 1/Metric 2	Low	Med	High
Low	holdback	re-test	implement
Med	re-test	implement	implement
High	implement	implement	implement

Scope of Test

Now we are getting to the part of the test spec that is more tactical in nature. This should really be used as a team to clarify exactly what is happening.

It’s often necessary to declare some things in scope or out of scope. This might seem like it's covered by the context or the hypotheses, but you’ll often find opportunities that are adjacent or similar to what’s covered by the test. It can be helpful to clarify the opportunity, by defining what must be done to accomplish for the test. Almost equally important is defining what is out of scope for the test. It is often useful to create some “straw-men” that describe what you are not trying to test. As you do this, you might circle back to your hypotheses (adding some as things that are nice-to-have are declared in scope for example).

Must-Do	**Nice-To-Have**	**Out-of-Scope**
- Some things are necessary, to address the hypotheses above	- Some things are close enough that they would be good to test if ready, but that we do not need to include to move forward	- Somethings we definitely will not get to in this test

For example: if you are sending emails to get new users to sign up of a product you might have the following examples:

- Must-do:
  - Create an email for new users.
  - Collect email addresses for potential users
- Nice to Have:
  - Test some different subject lines in the email
  - Test sending 1 email per week vs. 2 email per week
- Out of Scope:
  - Sending push notifications or social media ads
  - Design a new different home page for users who come from the email vs when searching on the web.

Measurement

Primary Metrics

What are the most important metrics for your test? These metrics are generally ones that go directly into the decision framework listed above. Ideally, there are only one or two. If there are more, you might want to declare a clear hierarchy amongst them as they go into you decision framework

- **Sometimes the metric will be obvious**. For example: it might be as simple at the total number of sales during a test of offering a discount
- Other times, there is **ambiguity in the metric**. For example, a credit card company might be interested in understanding how rewards programs increase the lifetime value of a customer (the total revenue they will make over the years owning the card). But, you aren't going to run an experiment long-enough to observe the full lifetime value of most members (remember, people hold on to credit-cards for years)
  - You might have to look for shorter term proxies such as:
    - Total revenue during the test
    - % increase in transactions during the test
    - # of people that switch to the new rewards program when offered to them

- Being specific in your hypothesis and your decision framework can help you understand which metric is most important
- Other times, additional research is necessary to understand what proxies tend to correlate with the true outcome (either in previous experiments, observational data ect).

There is often a **bias/variance** tradeoff in your choice in metric:

- A low variance metric is one that is "**sensitive**", it moves quickly in response to a change. In the example above, we would expect that the number of users that opt-into new rewards programs changes pretty quickly after being offered. This can be good, because it is often easy to see a test result in a pretty short amount of time, with a small sample size.
- However, frequently, you will find that the most-sensitive metrics are also **biased**, in that they can deviate from the true outcome of interest (or don't predict it well). For example, you might find that people who opt-in to the new reward program also end up using the credit card less and you make less money!

Often there is not a perfect answer on how to weigh these two concerns. It can be a whole separate line of research not addressed here to create a very credible primary metric, but it also left to the judgement of the team to weigh these concerns.

- There may be **practical constraints t**hat drive the choice in metrics such as
  - What data can be observed?
  - How long can we run the test for? When do you need to make a decision?
- There can also be **research** that informs this
  - What is the variance of a metric in observational data?
  - How well does a change in (proxy) predict a change in (true outcome) from other tests?
- Finally, there may just be some **domain expertise/historical knowledge** helps define the metric
  - A decision may have been made in the past on a metric, that turned out to be considered a failure. This motivates using a different metric in the future.

Secondary Metrics

These are metrics that are mainly designed to tell us if the test is working as intended. We should be able to hypothesize good secondary metrics

- For example, if you think you are going to increase revenue for a food delivery app by promoting certain restaurants on the home page, we would probably expect the number of orders from the promoted restaurants to increase. If the number of orders from the restaurants decreased, even if revenue went up, we might be skeptical, and wonder if we have a false positive.

Duration

There are often two consideration in how long to run a test for

- **Effect sizes:** Often, an effect size will increase over the course of a test. For example, the longer you run an ad campaign for, the more people will sign up during the campaign. In this case, running a test for longer will increase your ability to get a read.
- **Bias:** Often running a test for a short period of time is not a good example of what happens over the long-term. Running longer tests can prevent these forms of bias.
    a. If you roll-out a new rewards program for a credit card, people might spend extra at places that are rewarded highly when it's top of mind, but might revert back to their previous behavior as the new program becomes less salient (a **novelty effect)**.
    b. **Seasonality** often matters a lot too (if you have rewards for theme parks, you might want to make sure you measure both the summer season when people travel, and fall when families are back in school)
    c. Finally, a short test might get people to do something they would do anyway a little bit sooner (they might make a big purchase now to get the rewards, when they would have made that purchase two weeks from now if not offered the reward). We call this **pull-forward.**

The first of these issues relates to the next section, statistical power. The second one often relates to the bias-variance tradeoff in the choice of metrics. It often requires good judgement to decide how long to run the test for, as there is no magic answer.

Regardless, it is often necessary to declare in advance either exactly how long a test will run for, or when you know to stop. For example, the recent Covid-19 vaccine trials were allowed to be evaluated once a certain number of people in the test were diagnosed with Covid. If you do not declare this advance, you might end up stopping when the results look positive for some particular reason that may or may not be valid.

Sample size:

It is often necessary to decide on the sample size of a test in advance. The sample size is usually driven by statistical power: that is the ability to observe an effect (rejecting the null hypothesis) when the effect exists. There are calculators to help find the sample size, but here are some considerations:

1. Your power is often driven by the sensitivity of your metric. The more sensitive your metric, the less sample size you need as you can reasonably expect to see a big change during your test.
2. You often need to “assume” an effect size. This can be done two ways:
    a. Look at **historical similar tests** to assume a reasonable effectize, and calculate power from there. For example, if you are testing a new subject-line in an email, you might now know how good this particular subject line is, but may have tested other subject-lines before and have a good guess about how the change affects your metric

    b. Think about what a **material effect.** What is the smallest effect size that you would want to know about. You don't need to be powered for smaller effect sizes, but should know about anything that is material

3. Some lucky organizations (like Google) have so many users that it's pretty easy to imagine that obtaining sample size is trivial (over just a few days, billions of people will visit their site, and we can measure any relevant change). Most of the time though, you will face a tradeoff in how many cells you can test, how long you run the test for and how powered you are. Like so many other points, it requires good judgement to balance.

Cell Design

A group of users who are all exposed to the same experience. Part of defining your test is being explicit about the cells (or experiences) that we create.

Cell #	Cell Name	Cell Description
1	Control	There is often a control cell in an experiment. This is usually the primary thing other cells will be compared against. 1. Frequently, this is the “current production” experience, or represents the status-quo as it is today. 2. In some cases, this is a removal of anything related to treatment. For example, if we wish to measure the effect of an ad campaign, our control may be not sending any ads at all. 3. Sometimes the “current production” may not be very well defined. For example, perhaps 1 person currently makes a decision about when to send emails using their judgement. For the purposes of an experiment, it is often good to get specific about what the treatment will be during the test, even if it is a stylized version of what actually happens.
2	Secondary Control	In some cases our treatment may generate a “side-effect” that is not part of the primary hypothesis, but we want to control for. ● For example, often adding a new element to a digital product can increase the time to load a page. ● It can be worth decoupling the primary effect of the intervention, from that add on, in this case by artificially slowing the load time of the control experience with no other change
3 +	Treatments...	Cells that represent the intervention implied by your hypothesis. - In these cells it is often good to be specific, by stating “control cell + change 1 + change 2 …”

Sometimes “a change” is very technical. It is on the team to do a good job summarizing the set of things that have to happen to mean the change.

- For example, the change might be a change in terms for a credit card rewards program.
    - This might include changing the % reward for different products, an email to tell users about the change, and change in the landing site where the user redeems the rewards.

The test spec should be as specific as necessary about this bundle of changes to make sure everyone is aligned on what needs to happen. Often this will require long sections describing the treatment, or a secondary document with all the specifics. This may go beyond the scope of a scientist, but it's often a good idea to check in on this, because you might find something odd or that requires a secondary control.