**7 A/B Testing Questions and Answers in Data Science Interviews**

Common Pitfalls and Solutions Running A/B Tests

A/B tests, a.k.a controlled experiments, are used widely in industry to make product launch decisions. It allows tech companies to evaluate a product/feature with a subset of users to infer how the product may be received by all users. Data scientists are at the forefront of the A/B testing process, and A/B testing is considered one of a data scientist’s core competencies. **Data science interviews reflect this reality**. Interviewers routinely ask candidates A/B testing questions along with business case questions (a.k.a metric questions, product sense questions) to evaluate a candidate’s product knowledge and ability to drive the A/B testing process.

In this article, we’ll take an **interview-driven** approach by linking some of the most commonly asked interview questions to different components of A/B testing, including selecting ideas for testing, designing A/B tests, evaluating test results, and making ship or no ship decisions. Specifically, we’ll cover **7** most commonly asked interview questions and answers.

Before you start reading, if you are a video person, feel free to check out this [YouTube video](https://youtu.be/X8u6kr4fxXc) for an abbreviated version of this post.

**Table of Contents**

1. [Before a Test — Not Every Idea Is Worth Testing](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#306e)
2. [Designing A/B Tests](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#9339)

* [How long to run a test?](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#5072)
* [Interference between control and treatment groups](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#22a0)
* [How to deal with interference?](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#555e)

3. [Analyzing Results](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#53bf)

* [Novelty and Primacy Effects](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#e963)
* [Multiple testing problem](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#4848)

4. [Making Decisions](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#3d12)

5. [Resources](https://towardsdatascience.com/7-a-b-testing-questions-and-answers-in-data-science-interviews-eee6428a8b63#0943)

**Before a Test — Not Every Idea Is Worth Testing**

An A/B test is a powerful tool but not every idea is selected by running a test. Some ideas can be expensive to test, and companies in early phases may have resource constraints, so it’s **not practical to run a test for every single idea**. Thus, we want to first select which ideas are worth testing, particularly when people have different opinions and ideas on improving a product, and there are many ideas from which to choose. For example, a UX designer may suggest changing some UI elements, a product manager may propose simplifying the checkout flow, an engineer may recommend optimizing a back-end algorithm, etc. In times like this, stakeholders rely on data scientists to drive data-informed decision making. A sample interview question is:

There are a few ideas to increase conversion on an e-commerce website, such as enabling multiple-items checkout (currently users can check out one item at the time), allowing non-registered users to checkout, changing the size and color of the ‘Purchase’ button, etc., how do you select which idea to invest in?

One way to evaluate the value of different ideas is to conduct **quantitative analysis** using historical data to obtain the **opportunity sizing** of each idea. For example, before investing in multiple items checkout for an e-commerce website, obtain an upper-bound sizing of the impact by analyzing the number of multiple-items purchased per user. If only a very small percentage of users purchased more than one item, it may not be worth the effort to develop this feature. It’s more important to investigate users’ purchasing behaviors to understand why users do not purchase multiple items together. Is it because the selection of items is too small? Are the items too pricey, and they can only afford one? Is the checkout process too complicated, and they do not want to go through it again?

This kind of analysis provides directional insights on which idea is a good candidate for A/B testing. However, historical data only tells us how we’ve done in the past. It’s not able to predict the future accurately.

To get a comprehensive evaluation of each idea, we could conduct **qualitative analysis** with focus groups and surveys. Feedback gathered from focus groups (guided discussion with users or perceptive users) or questions from surveys provide more insights into users’ pain points and preferences. A combination of qualitative and qualitative analysis can help further the idea selection process.

**Designing A/B Tests**

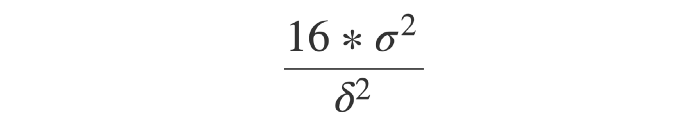
Once we select an idea to test, we need to decide how long we want to run a test and how to select the randomization unit. In this section, we’ll go through these questions one by one.

**How long to run a test?**

To decide the duration of a test, we need to obtain the sample size of a test, which requires three parameters. These parameters are:

* Type II error rate *β* or Power, because Power = *1 - β*. You know one of them, you know the other.
* Significance level *α*
* Minimum detectable effect

The rule of thumb is that sample size *n* approximately equals 16 (based on *α = 0.05* and *β = 0.8*) multiplied by *sample variance* divided by *δ* square, whereas *δ* is the difference between treatment and control:



This formula is from [Trustworthy Online Controlled Experiments](https://www.amazon.com/Trustworthy-Online-Controlled-Experiments-Practical/dp/1108724264) by Ron Kohavi, Diane Tang and Ya Xu

If you are interested in learning how we come up with the rule of thumb formula, check out [this video](https://youtu.be/JEAsoUrX6KQ) for a step by step walkthrough.

During the interview, you are not expected to explain how you came up with the formula, but you do want to explain how we obtain each parameter and how each parameter influences the sample size. For example, we need more samples if the sample variance is larger, and we need fewer samples if the delta is larger.

Sample variance can be obtained from the existing data, but how do we estimate *δ*, i.e the difference between treatment and control?

Actually, we don’t know this before we run an experiment, and this is where we use the last parameter:**the minimum detectable effect**. It is the smallest difference that would matter in practice. For example, we may consider a 0.1% increase in revenue as the minimum detectable effect. In reality, this value is discussed and decided by multiple stakeholders.

Once we know the sample size, we can obtain the number of days to run the experiment by dividing the sample size by the number of users in each group. If the number is less than a week, we should run the experiment for at least seven days to capture the weekly pattern. It is typically recommended to run it for two weeks. When it comes to collecting data for a test, **more is almost always better than not enough**.

**Interference between control and treatment groups**

Normally we split the control and treatment groups by randomly selecting users and assigning each user to either the control or treatment group. We expect each user is independent and no interference between control and treatment groups. However, sometimes **this independence assumption does not hold**. This may happen when testing social networks, such as Facebook, Linkedin, and Twitter, or two-sided markets such as Uber, Lyft, and Airbnb. A sample interview question is:

Company X has tested a new feature with the goal to increase the number of posts created per user. They assigned each user randomly to either the control or treatment group. The Test won by 1% in terms of the number of posts. What do you expect to happen after the new feature is launched to all users? Will it be the same as 1%, if not, would it be more or less? (assume there’s no novelty effect)

The answer is that we will see a value larger than 1%. Here’s why.

In **social networks** (e.g. Facebook, Linkedin, and Twitter), users’ behavior is likely impacted by that of people in their social circles. A user tends to use a feature or a product if people in their network, such as friends and family, use it. That is called a **network effect**. Therefore if we use “user” as the randomization unit and the treatment has an impact on users, the effect may spill over to the control group, meaning people’s behaviors in the control group are influenced by those in the treatment group. In that case, the difference between control and treatment groups **underestimates** the real benefit of the treatment effect. For the interview question, it will be more than 1%.

For **two-sided markets** (e.g. Uber, Lyft, ebay, and Airbnb): interference between control and treatment groups can also lead to biased estimates of treatment effect. It is mainly because resources are shared among control and treatment groups, meaning that control and treatment groups will compete for the same resources. For example, if we have a new product that attracts more drivers in the treatment group, fewer drivers will be available in the control group. Thus, we’ll not be able to estimate the treatment effect accurately. Unlike social networks where the treatment effect underestimates the real benefit of a new product, in two-sided markets, the treatment effect **overestimates** the actual effect.

**How to deal with interference?**

Now that we know why interference between control and treatment can cause the post-launch effect to behave differently than the treatment effect, it leads us to the next question: **how do we design the test to prevent the spillover between control and treatment?** A sample interview question is:

We are launching a new feature that provides coupons to our riders. The goal is to increase the number of rides by decreasing the price for each ride. Outline a testing strategy to evaluate the effect of the new feature.

There are many ways to tackle the spillover between groups and the main goal is to **isolate users** in the control and treatment groups. Below are a few commonly used solutions, each applies in different scenarios, and all of them have limitations. In practice, we want to choose the method that works the best under certain conditions, and we could also combine multiple methods to get reliable results.

**Social networks:**

* One way to ensure isolation is to create **network clusters** to represent groups of users who are more likely to interact with people within the group than people outside of the group. Once we have those clusters, we could split them into control and treatment groups. Check out [this paper](https://engineering.linkedin.com/blog/2019/06/detecting-interference--an-a-b-test-of-a-b-tests) for more details on this approach.
* **Ego-cluster** randomization. The idea originated from Linkedin. A cluster is composed of an “ego” (a focal individual), and her “alters” (the individuals she is immediately connected to). It focuses on measuring the one-out network effect, meaning the effect of a user’s immediate connection’s treatment on that user, then each user either has the feature or does not, and no complicated interactions between users are needed. [This paper](https://arxiv.org/abs/1903.08755) explains this method in detail.

**Two-sided markets:**

* **Geo-based** randomization. Instead of splitting by users, we could split by geo-locations. For example, we could have the New York Metropolitan area in the control group and the San Francisco Bay Area in the treatment group. This will allow us to isolate users in each group, but the pitfall is that there will be a larger variance because each market is unique in certain ways through things such as the customer’s behavior, competitors, etc.
* Another method, though used less commonly, is **time-based** randomization. Basically, we select a random time, for example, a day of a week, and assign all users to either the control or treatment group. It works when the treatment effect only lasts for a short amount of time, such as when testing a new surge price algorithm performs better. It does not work when the treatment effect takes a long time to be effective, such as a referral program. It can take some time for a user to refer to his or her friends.

**Analyzing Results**

**Novelty and Primacy Effects**

When there’s a change in the product, people react to it differently. Some are used to the way a product works and are reluctant to change. This is called the **primacy effect** or change aversion. Others may welcome changes, and a new feature attracts them to use the product more. This is called the **novelty effect**. However, both effects will not last long as people’s behavior will stabilize after a certain amount of time. If an A/B test has a larger or smaller initial effect, it’s probably due to novel or primacy effects. This is a common problem in practice, and many interview questions are about this topic. A sample interview question is:

We ran an A/B test on a new feature and the test won, so we launched the change to all users. However, after launching the feature for a week, we found that the treatment effect quickly declined. What is happening?

The answer is the novelty effect. Over time, as the novelty wears off, repeat usage will be decreased so we observe a declining treatment effect.

Now you understand both novelty and primacy effects, **how do we address the potential issues**? This is a typical follow-up question during interviews.

One way to deal with such effects is to completely rule out the possibility of those effects. We could run tests only on first-time users because the novelty effect and primacy effect obviously doesn’t affect such users. If we already have a test running and we want to analyze if there’s a novelty or primacy effect, we could 1) compare new users’ results in the control group to those in the treatment group to evaluate novelty effect 2) compare first-time users’ results with existing users’ results in the treatment group to get an actual estimate of the impact of the novelty or primacy effect.

**Multiple testing problem**

In the simplest form of an A/B test, there are two variants: Control (A) and treatment (B). Sometimes, we run a test with multiple variants to see which one is the best amongst all the features. It can happen when we want to test multiple colors of a button or test different home pages. Then we’ll have more than one treatment group. In this case, we should not simply use the same significance level of 0.05 to decide whether the test is significant because we are dealing with more than 2 variants, and the probability of false discoveries increases. For example, if we have 3 treatment groups to compare with the control group, what is the chance of observing at least 1 false positive (assume our significance level is 0.05)?

We could get the probability that there is no false positives (assuming the groups are independent),

Pr(FP = 0) = 0.95 \* 0.95 \* 0.95 = 0.857

then obtain the probability that there’s at least 1 false positive

Pr(FP >= 1) = 1 — Pr(FP = 0) = 0.143

With only 3 treatment groups (4 variants), the probability of a false positive (or Type I error) is over 14%. This is called the “**multiple testing**” problem. A sample interview question is

We are running a test with 10 variants, trying different versions of our landing page. One treatment wins and the p-value is less than .05. Would you make the change?

The answer is no because of the multiple testing problem. There are several ways to approach it. One commonly used method is **Bonferroni correction**. It divides the significance level 0.05 by the number of tests. For the interview question, since we are measuring 10 tests, then the significance level for the test should be 0.05 divided by 10 which is 0.005. Basically, we only claim a test if significant if it shows a p-value of less than 0.005. The drawback of Bonferroni correction is that it tends to be too conservative.

Another method is to control the **false discovery rate** (FDR):

FDR = E[# of false positive / # of rejections]

It measures out of all of the rejections of the null hypothesis, that is, all the metrics that you declare to have a statistically significant difference. How many of them had a real difference as opposed to how many were false positives. This only makes sense if you have **a huge number of metrics**, say hundreds. Suppose we have 200 metrics and cap FDR at 0.05. This means we’re okay with seeing false positives 5 of the time. We will observe at least 10 false positive in those 200 metrics every time.

**Making Decisions**

Ideally, we see practically significant treatment results, and we could consider launching the feature to all users. But sometimes, we see **contradicting results**, such as one metric goes up while another one goes down, so we need to make a win-lost tradeoff. A sample interview question is:

After running a test, you see the desired metric, such as the click-through rate is going up while the number of impressions is decreasing. How would you make a decision?

In reality, it can be very involved to make product launch decisions because various factors are taken into consideration, such as the complexity of implementation, project management effort, customer support cost, maintenance cost, opportunity cost, etc.

During interviews, we could provide a simplified version of the solution, focusing on the **current objective** of the experiment. Is it to maximize engagement, retention, revenue, or something else? Also, we want to quantify the negative impact, i.e. the negative shift in a non-goal metric, to help us make the decision. For instance, if revenue is the goal, we could choose it over maximizing engagement assuming the negative impact is acceptable.

**Resources**

Lastly, I’d like to recommend two resources for you to learn more about A/B testing.

* [Udacity’s free A/B testing course](https://www.udacity.com/course/ab-testing--ud257) covers all the fundamentals of A/B testing. Kelly Peng has a[great post](https://towardsdatascience.com/a-summary-of-udacity-a-b-testing-course-9ecc32dedbb1) summarizing the content of the course.
* [Trustworthy online controlled experiments — A practical guide to A/B testing](https://www.amazon.com/Trustworthy-Online-Controlled-Experiments-Practical/dp/1108724264) by Ron Kohavi, Diane Tang, and Ya Xu. It has in-depth knowledge on how to run A/B tests in industry, the potential pitfalls, and solutions. It contains a lot of useful stuff, so I actually plan to write a post to summarize the content of the book. Stay tuned if you are interested!