

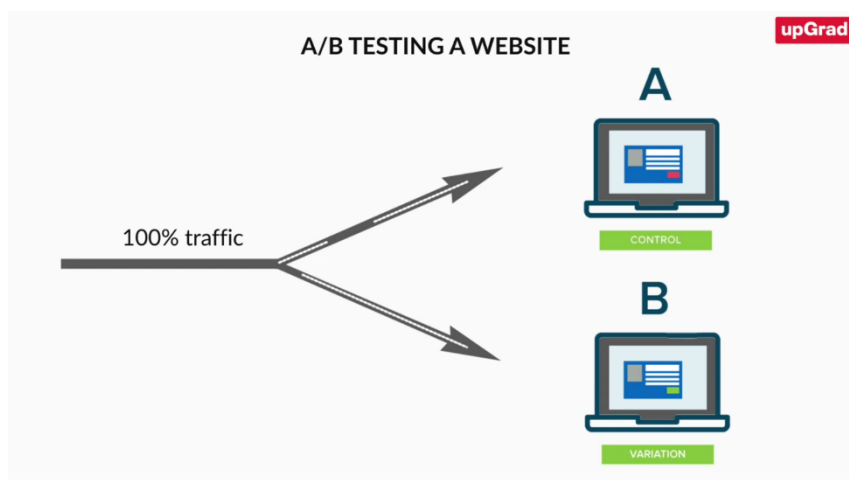
Transcription

A/B testing



Hi there. In the previous session, you became familiar with part analytics and learned how to track the important metrics. Now a question arises in my mind about the relation between aesthetics and functionality of the website.

Now we all know that a well-designed website leads to better user engagement, so aesthetics are an important aspect, but there are so many ways of arranging elements in the website. This might lead to different opinions. In such a scenario, how do we make rational decisions about things like say the button or colours of various elements or what text should appear on the website?

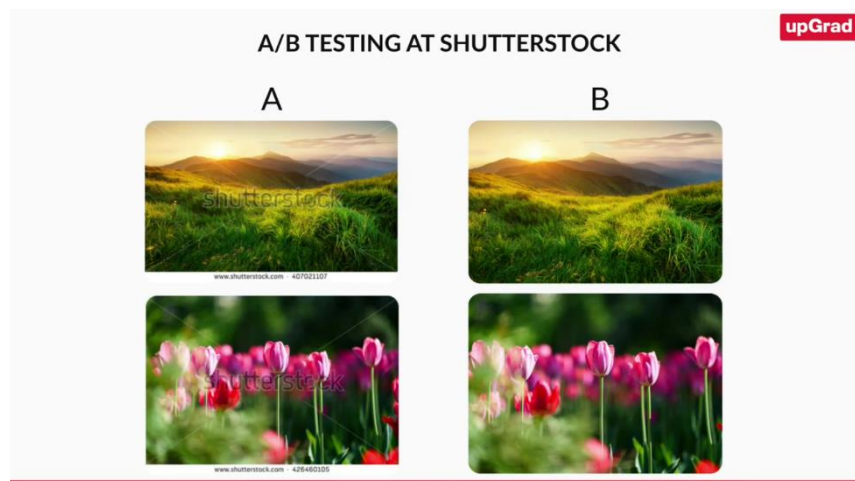


You can use AB testing. AB testing provides you a way to test two different versions of the same element. Let's see how exactly this works. Let's say you are working on a website. Suppose A is the existing design of your website, this design is called the control. B is the newly designed version, which is called the variation.

To conduct the AB test, you would have to divide the traffic between the control and the variation, that is the A and the B, and measure the performance of each design using the matrices which are important to your business. At the end of the test, the data collected would help you select the better performing version for your website.

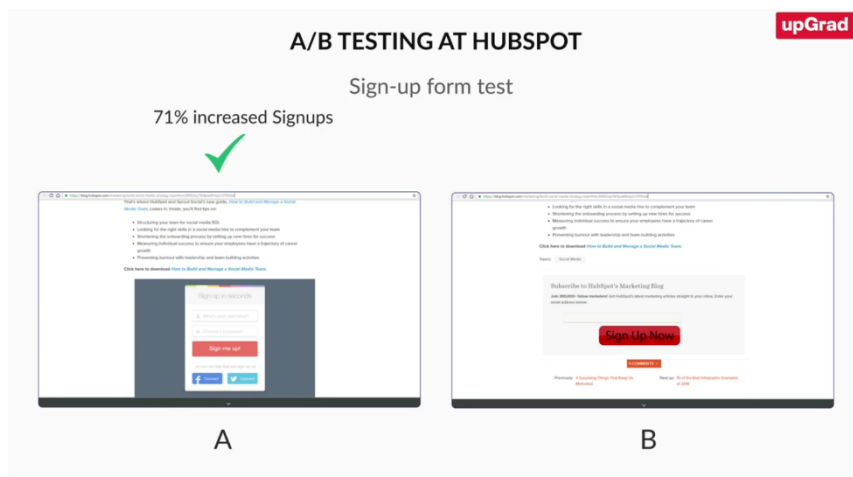


One of the great examples of AB testing is one done by Shutterstock. It's a marketplace for images, creatives and illustration. Now Shutterstock sees about three downloads per second, so they have a ton of data about their users and can run multiple AB tests with statistical significance. The company run AB test for multiple things like text, colour of links, pricing and search algorithms.



One such test they ran was related to the watermark shown on the images. Company had the hypothesis that removing the watermark would lead to increased conversions because there would be far less visual clutter on the page.

But when AB test for the same was run, two groups were shown two different pages, one with watermark on images and other without the watermark. The data collected disproved the hypothesis as the company learned that the user who saw the images without watermark had a lower conversion rate.



Another great example of AB testing is a sign of form test done by a company called HubSpot, which does internal marketing. They tested it on their blog. In one variation, the form was embedded in the poster itself. That is, it was an in-line CTA. The other variation was a link to a landing page with the form.

The first variation where the form was on the same page as the blog post performed better than the second one by 71%. So, you can see how AB can resolve different opinions within the team.



AB testing is a better way to arrive at conclusions. We can calibrate values of new elements and see what really works. It allows me to test the hypothesis with the user feedback. But beyond this, how does this help a PM in achieving business goals?




GOALS OF A/B TESTING

Type of business	Business goal
E-commerce	Increased sales
SaaS	Freemium
Blog	Signups

Different types of websites have different business goals. For example, an eCommerce website wants users to buy products at the end of the day, a SAS provider wants user to sign up for trial and then convert them to paid user. A blog or a news website would want the user to sign up for a mailing list or a premium version.

In all these cases, the end goal is the conversion of user. AB testing helps you find out what you could change in the current product to help you increase this conversion rate.



ATTRIBUTES OF AN A/B TEST

- 01 Removes guesswork
- 02 Tests various hypotheses
- 03 Uses a data-driven methodology
- 04 Enables to increase conversions

Running AB tests removes guesswork and provides an opportunity to test various hypothesis regarding the users using data driven methodology.

In a team working on a product, there may be different opinions regarding the design of the product. Running AB test helps you resolve this conflict by choosing the versions backed by data.

AB test helps you make more out of your current traffic. Now increasing traffic itself is a lot more expensive, a lot more difficult, and hence, increasing the conversion of current traffic is a cheaper and faster option.

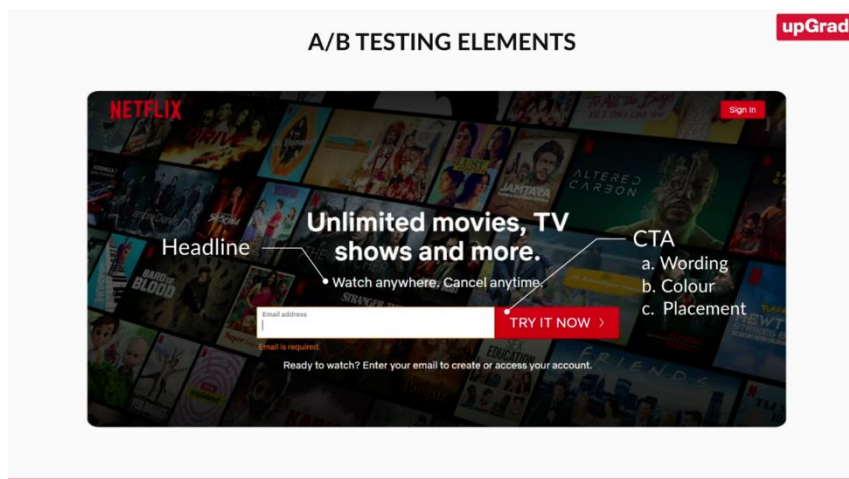


So, in this video we learned what AB testing is and you saw how AB testing plays an important role in improving conversion, that's helping a PM reach the business goals. We now know that the user data authenticates the credibility of AB testing. And since it leads to better conversions, it's also cheaper and a faster option. Now in the next video, we'll learn more about performing AB testing.

Now you know that by AB testing we can decide which elements can lead to better conversions. Like me, you must be wondering, but what are these various elements that can be tested?



Your choice of what to test will depend on the goal of your test. For example, your goal could be to increase the number of signups. In this case, you could test elements like length of sign up form, the type of the fields in the form. The goal of the AB testing in this case is to figure out what prevents users from signing up.

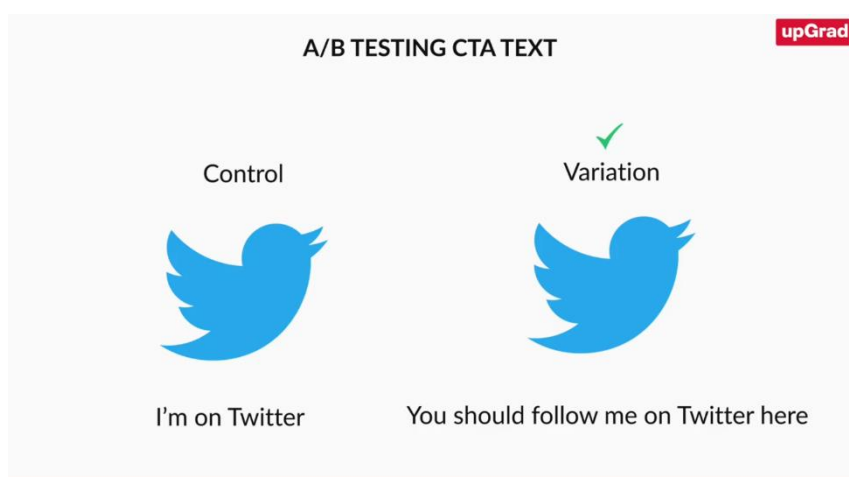


Every AB test you carry out would be unique, but there are certain elements which usually come under the purview of AB tests. Let's list a few of these elements.

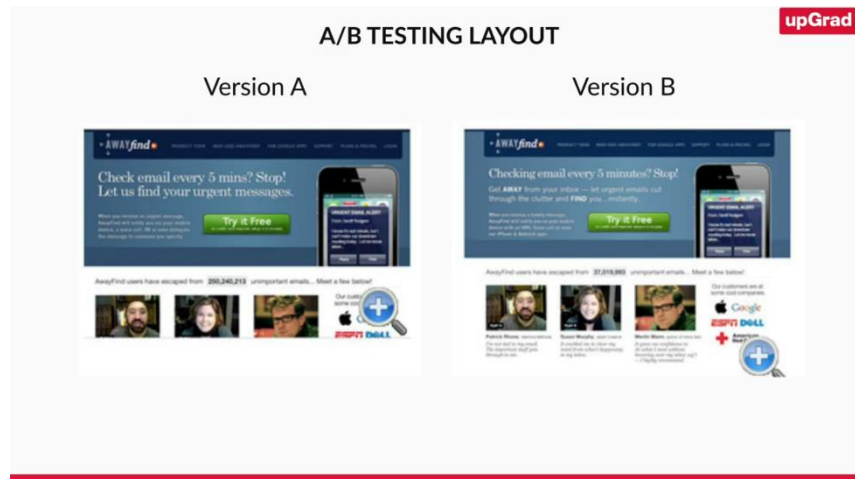
- The first one is called call to action or CTA, for example, a button. What tested here is the wording, colour and the placement.
- The second one is the headline or the product description.
- Third one is the form splint and the type of fields it would contain.
- The fourth one is the layout or more broadly the style of the website.
- Next comes the product pricing and the promotional offers.
- Then the images on the landing page and the product pages.
- Also, the amount of text on the page, whether the copy is too short or long is also usually tested.

Now that you have the broad overview, let's now look at these in detail. Let's start with text. AB testing, the copy of your text can help you increase conversions. You can try out changes in the headings, paragraph, CTA buttons and the text of form fields.

A good example of text change leading to increased conversion is the one by Dustin Curtis, a well-known web designer and founder of Svbtile, a blogging platform.

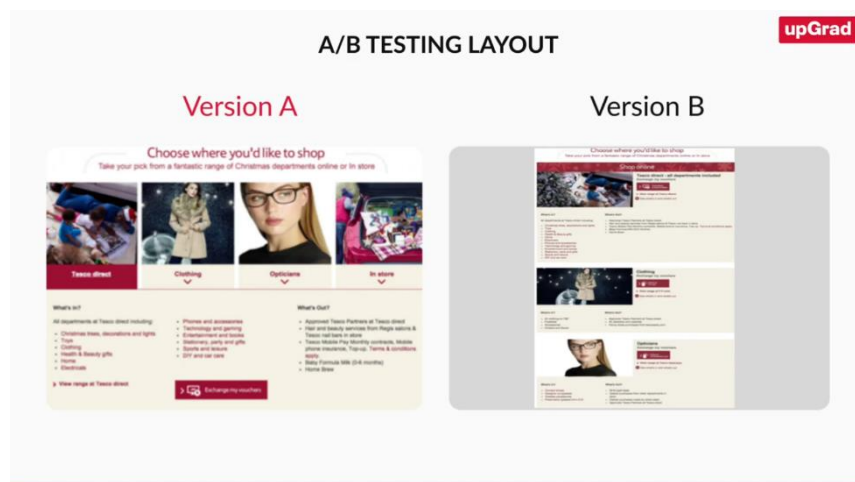


He tested two different types of CTA text on his blog homepage. The control version was called, I'm on Twitter, and the variation was, you should follow me on Twitter here. After running the experiment, he found that the variation worked 173% better than the control text.



Let's check out another example. On the screen, you can see two versions of the same page. An AB test was carried out and it was found out that the version B increased signups by 38%. Now looking at the versions, you might think version is better design wise. By comparison, it has a clear headline, a supporting copy, is short and crisp. But in the AB test, version B performed better.

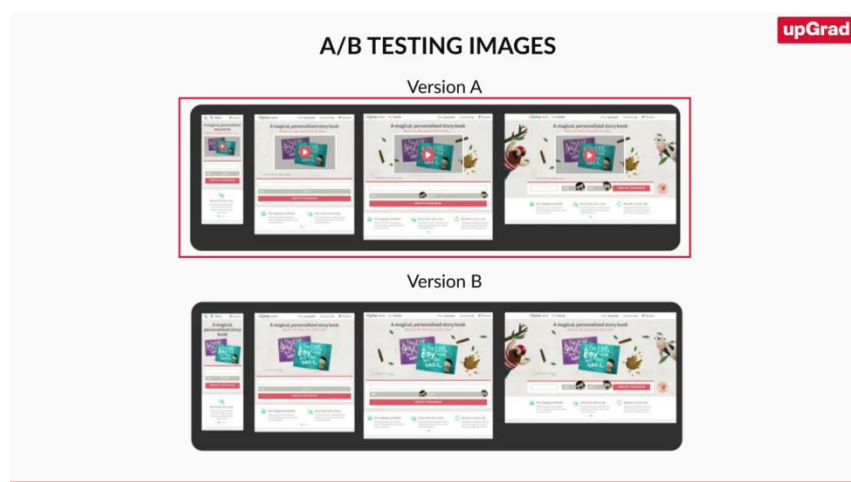
Why? Why does version B work? Simply because copy blocking is better. The headline is shorter. The subheading is designed to pick out some key features in board. It's not as pretty, but the information transferred to the user is more efficient because of the emphasis within the copy. The learning from this is that the clean design doesn't necessarily mean an effective one.



Next let's talk about the layout of your website. Now using AB testing, you can also test various combination of different elements of your website.

You can see the example of this here. As you can see, version A has a horizontal layout with a CTA button at the bottom of the page. Well version B has a vertical layout with a CTA placed next to each image. Upon testing both of these pages, it was found that version A increased conversion rates higher than version B.

Let's talk about images now. You can test images using AB testing. For instance, you can test images, which images work from the group of images. You can test if placing an image on a page increases conversion in many such cases.



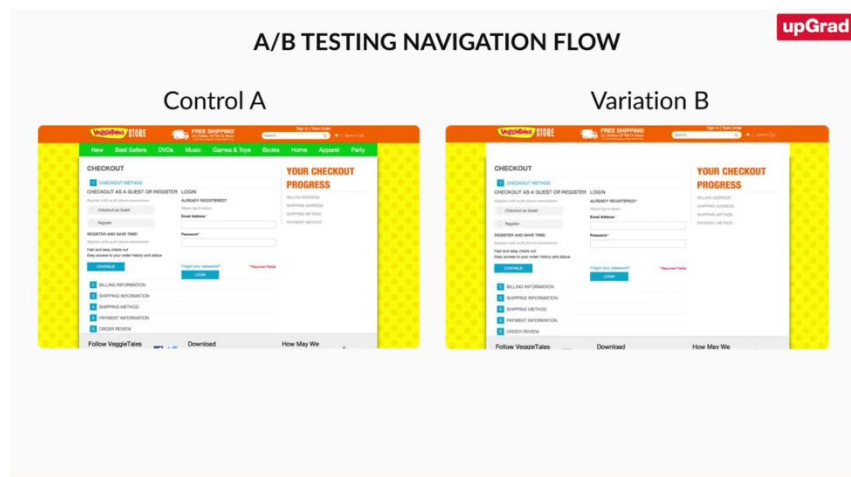
Let's see, an example of this. Here are two different versions of a homepage. Version A has a video prominently placed at the top of the page, whereas version B has an image in place of a video. Upon doing an AB test, it was found that version A with a video instead of an image had a higher conversion rate.

Similarly, you can also test for different placements of image on the web. Now, in the next experiment, you can see that version A has a static image on the page, while version B has an image slider. After running the image, it was found that version A had higher conversion than version B.

So, while I was at Goibibo, one of our major business was collecting leads. We experimented on a lead form in various different ways. Let's talk about what we did with images.

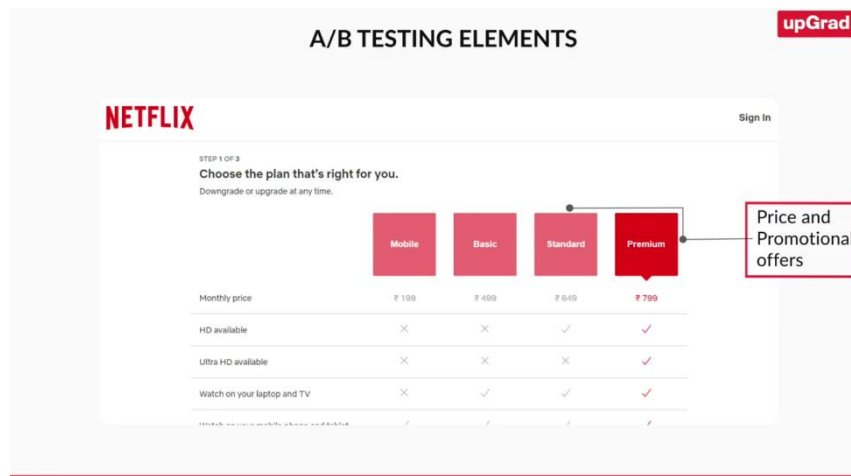
We realized that every time we had put an image of a notion, our honeymoon lead form performed better than if you would put a desert for instance. And a generate lead form with a woman as an image would always perform better than a generic hotel images on the lead form.

So, these are the things which you can only derive. You can have a hypothesis, but you can only derive once you start running AB test.



Next up, let's talk about navigation flow. AB test can also help you test multiple navigational flow between the different pages on your website. For example, one of the animated children TV series called VeggieTales ran an interesting experiment.

In one experiment, they removed the top navigation from their checkout pages. As you can see in the graphics, such a navigation panel, all the present in the control is absent in the variation. Once the customer reached the checkout page in the variation, they didn't have the option to navigate to other pages on the site. This change resulted in the 14% increase in revenue per visitor.



Further, you can also AB test promotions and different pricing tiers. You can test combination of pricing and also test including a promotion increases the conversion. Use of AB testing for pricing pages, this is actually very interesting.

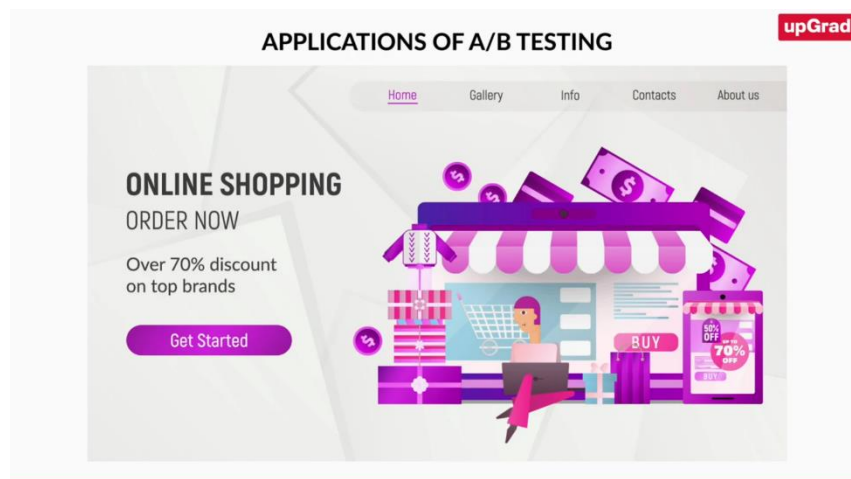
Let us use a real-life example. You are deciding between two vacation options, France and Germany. They're both great options, but each has their own advantages and disadvantages. A few days later, you find a deal of vacation to Germany with free meals. Then the decision became clear bye, bye Eiffel Tower and hello Germany with free breakfast.

Why did you make this choice? Well, because the inferior option of Germany without meals, you're drawn to the option of Germany with meals so much so that this is far more appealing than that of France. Such human behaviour is used by the companies while deciding their pricing structure.

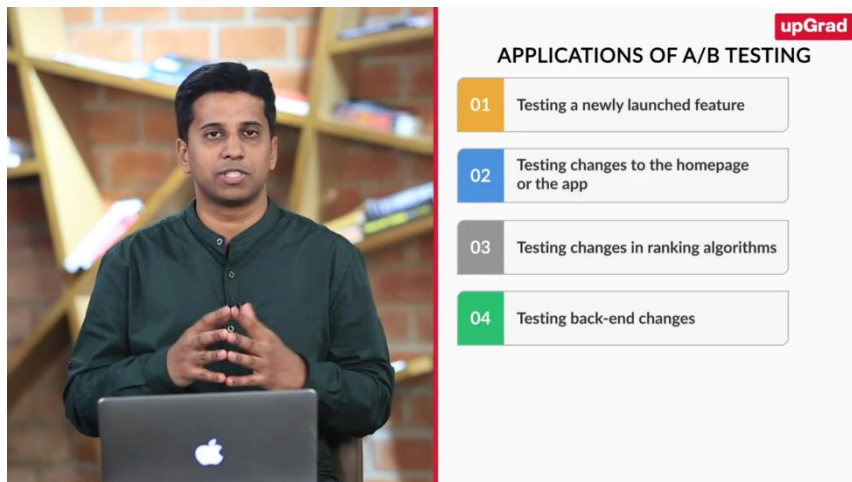
PRICING TIERS				
Control		Variation		
CONTROL	CONTROL	VARIATION	VARIATION	VARIATION
\$2.99 Monthly	\$29.99 Yearly	\$2.99 Monthly	\$24.99 Yearly	\$29.99 Yearly
<ul style="list-style-type: none"> Class 3 photographers Post production high quality images in one week Pay/month 	<ul style="list-style-type: none"> 2 free shoots/year Class 1 photographers Post production high quality images in 3 days 	<ul style="list-style-type: none"> Class 3 photographers Post production high quality images in one week Pay/month 	<ul style="list-style-type: none"> Class 2 photographers Post production high quality images in 1 week Indoor shoots only 	<ul style="list-style-type: none"> 2 free shoots/year Class 1 photographers Post production high quality images in 3 days
SUBSCRIBE	SUBSCRIBE	SUBSCRIBE	SUBSCRIBE	SUBSCRIBE

For example, let's look at an example of Smart Shoot, which AB tested two versions of their pricing pages. The control variation had two options, monthly or yearly with the same number of features under both the plans and users ended up using monthly options more. The variation B had an inferior product to the yearly plan as the company wanted users to buy that.

So, they created another tier of less yearly price, but this version had only core feature, not the extra one. The result was 233% increase in conversion, and 86% users ended up choosing the yearly option, while only 14% chose the monthly option.



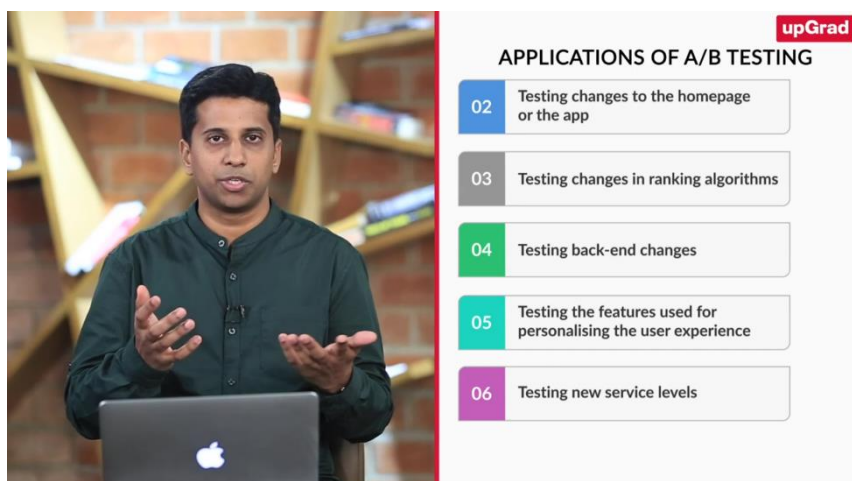
So, in what situations can you do AB testing? Is it always about some homepage banner, or is it some other, let's say call to action? Well, you can use AB testing in a lot of situations.



You can do it whenever you are doing a new feature launch and you want to test whether this feature will help you or not. Or if you're doing some changes on the homepage on your app, let's say you're changing some colours or some size of the different options or the layout and so on. Or other user experience, some changes in the whole app and flow in navigation.

And it would also have things which are not so apprehend. You can have things like ranking algorithms, which run behind the scene. So, for example, if you're searching on the Flipkart app, there could be a lot of different algorithms running behind to fetch you the right results.

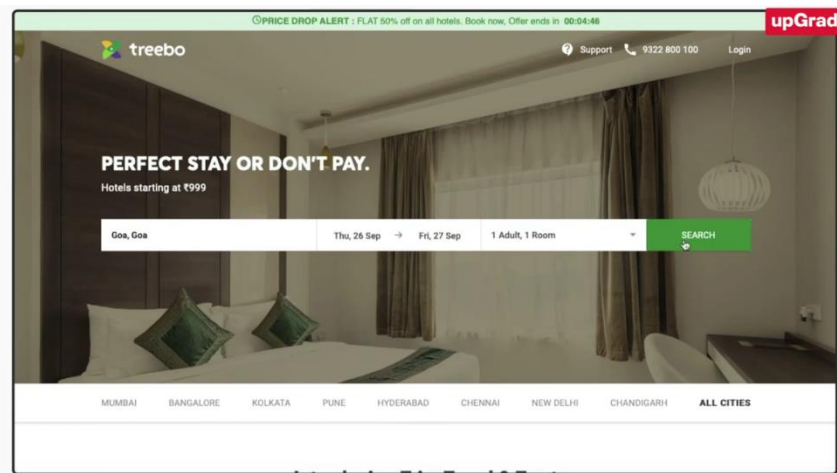
So, you could compare algorithm A versus algorithm B as well. Or you could have maybe the way the review is ranked, maybe that can be changed. So, you can have different review ranking algorithms for the reviews model. Or you could be just trying out different pricing structures. Maybe in one structure, you show the offer, in other structure, you don't show the offer, or maybe just you're trying out different pricing algorithms in the background.



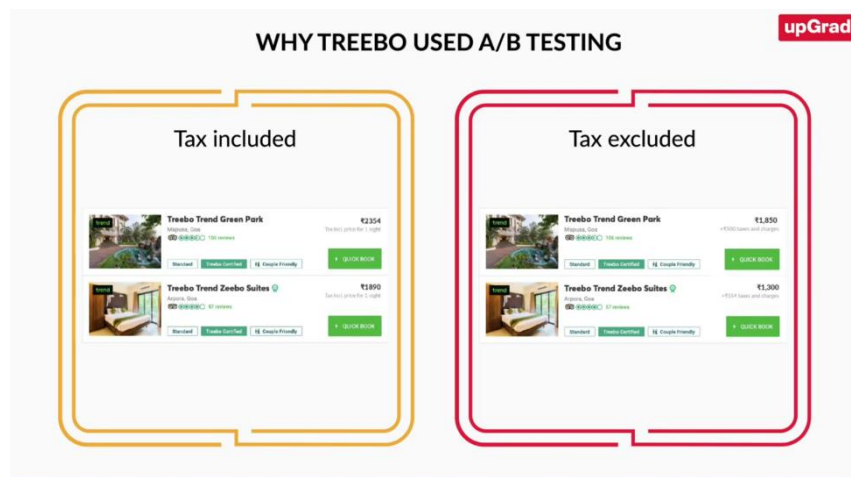
Or it could be some things which are not even so apprehend. It could be something which is say, some backend change, which helps you load some content faster, faster loading of the site. Maybe you want to test that. Or it could be something around personalizing the experience for the user.

Or it could also be something around say you promising a different delivery, time delivery estimate to the customer on your eCommerce site, and many more. So, any place or almost anything on your website, a property or some algorithm or anything which can affect your user behaviour can be tested through an AB testing setup.

To understand the importance of AB testing, to begin with, let's take the example of Treebo, and why they felt the need to invest heavily into their own AB platform.



Treebo, as many of you may know, is among the top budget hotels in India. And they have a lot of listings on their site. When someone comes in, the person can see the price against each hotel before they go and make the booking.



Now, Treebo had a couple of options here on the pricing. They could first show the full price of the hotel, all taxes inclusive, the final price, which the customer will eventually pay, or they could show a lower price without any other taxes applied which attracts more customers. They go further. And finally, on the booking page, when they have to make the payment, that's where they show the full price inclusive of all taxes.

So, there are two options that Treebo had. Each of them has its own merits. And actually B, the second option is what a lot of players do. So, Treebo wanted to evaluate this, Treebo was actually using the first option, which they were showing the full price inclusive of all taxes. And then one day, they decided let's switch to the other option because most of the industry seems to be doing that, show pre-tax prices.

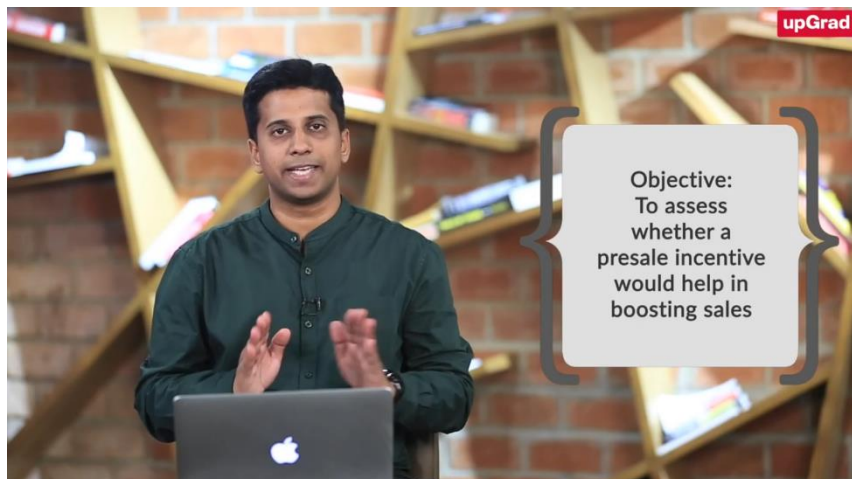
So, that's what they wanted to do. And on one Friday night, they switched completely to a pre-tax price and it backfired badly. Not only that, that they were not able to monitor in real time or even stop this, and it hit their conversions very badly, and resulted in one of the worst weekends ever in terms of their sales.



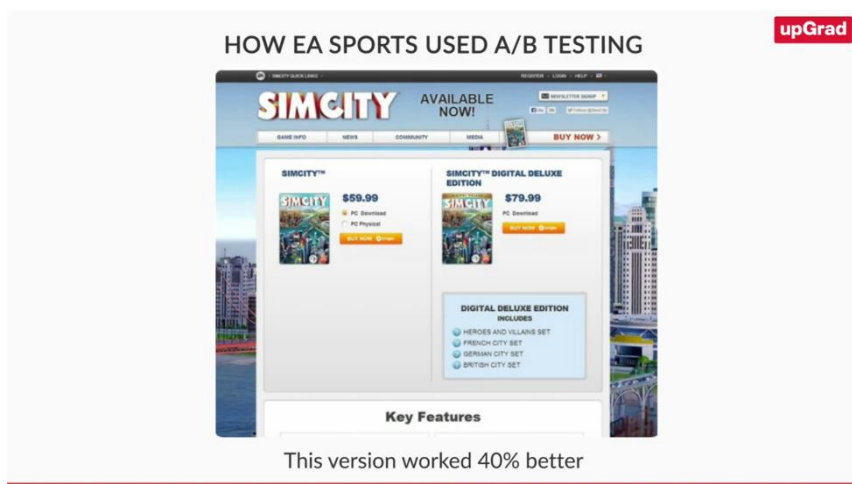
So, the question they now had was why did it fail and how is it working so well for so many other players. So, a detailed root cause analysis followed. And when they tried to dig deeper, they realized there were three issues that were there before they went into this.

1. First, they exposed this new change to all the traffic at once. So, they never tried, let's see if it works for 20% audience or 10% audience. How they react, they didn't do that. They straight away went and just exposed the entire audience to this change.
2. Second that they had no easy way of switching this off. They changed how the prices were being served completely and they just could not take it back immediately. In fact, they couldn't even monitor it in great detail for them to this action.
3. The third thing was that even after doing all of this, they were not sure whether the sales bombing, this was because of that price change, which they made or was it because of some other reason altogether? Because there could have been some other influencing factors as well.

So, all of these questions and there's no way to answer these questions. And that's when Treebo realized that they really have to invest in their own AB platform. And then they went ahead used an open source tool and then they made their own platform calling it Akkad Bakkad. And then from then on, they have a lot of learnings to share and it has helped them measure or analyse or test things so much better.



Now let's look at an example from the gaming industry, the example from electronic arts or EA games. And in this case, they wanted to assess whether giving an incentive to the customers, the presale incentive would help in boosting the sales. This is a fair hypothesis that giving an incentive to the customer would help boost the sales.



And in this case, they had two variants of the experience of the sales page. Variant one was where we have the usual sales call out, plus you have an extra bar talking about pre sales incentive, which is if you pre-order, you can unit now, you will get a \$20 discount on your next purchase. This was one.

The other experience didn't have any of this pre sales pitch. They just had straight go and buy brighter. So, this is one good example where your conventional wisdom and your usual hypothesis failed. In this case, the other variant without any incentive worked 40% better than the first one with the incentive.

So, in this case, maybe the fans, the Ebit fans just wanted to buy the game as quickly as possible. So, in this case, we saw that maybe a conventional wisdom doesn't really work that well, and sometimes it's best to see, try it out, and just see what the data tells us.



Before you start thinking about running an AB test, you first need to analyse the data already with you. Analysing this data will help you figure out a couple of things. The first one is, identify what you need to test and prioritize what should be tested first.

So, what should you analyse to figure out the above? Well, it depends a lot on the industry your product serves. For example, for e-commerce, it would be the web traffic, order history, customer support, tickets and drop offs.

Let's say in case of travel industry, you would measure number of room nights, number of transactions. What would be the room night is to transaction ratio, what would be the bounce rate? How are the users flowing? What would be the segmentation of users coming from SEO versus direct?

And of course, if you have a major customer support organization, you need to make sure that they are taking care of while you're running your AB test.



In this video, you learned that before performing an AB test, you have to analyse the available data. And that's all from this video. In the next video, we will understand how to formulate a hypothesis. See you then.

GOOD HYPOTHESIS FORMAT

upGrad

✓ ✓ ✓
If (variable), then (results) because (rationale)

So, now that you have decided what to test and where to test based on analysis of your data, the next step is to form a hypothesis for AB test. The data collected during the test will either prove or disprove your hypothesis. A good hypothesis would follow this format. If variable, then results because rationale. Let's look at what each of these elements actually means.

So, the variable refers to the single elements that that will be modified for the test. The result refers to the predicted outcome and could be something like more email signups or more clicks.

Finally, the rationale is what would lead to the increase in the results or what would be proved wrong if hypothesis is proved wrong.

Here is an example of a hypothesis in this format. If the call to action is changed to buy now, the conversion rates will increase because it is more relatable than shop now, which is the current CTO for example. But take care while applying this format because even with this, you can end up with good and a bad hypothesis.

GOOD VS BAD HYPOTHESIS

upGrad

Hypothesis 1: If the check out flow is reduced to fewer pages, the checkout completion rate will increase.

Hypothesis 2: If the navigation is removed from the checkout pages, the conversion rates will increase because the analytics have shown that people drop off while navigating between pages.

Let's look at some examples to understand the difference between good and a bad hypothesis. If you want to test a different navigation flow for the checkout on an eCommerce website like Flipkart, the possible hypothesis could be written as one, let's say if the checkout flow is reduced to fewer pages, the checkout completion rate will increase.

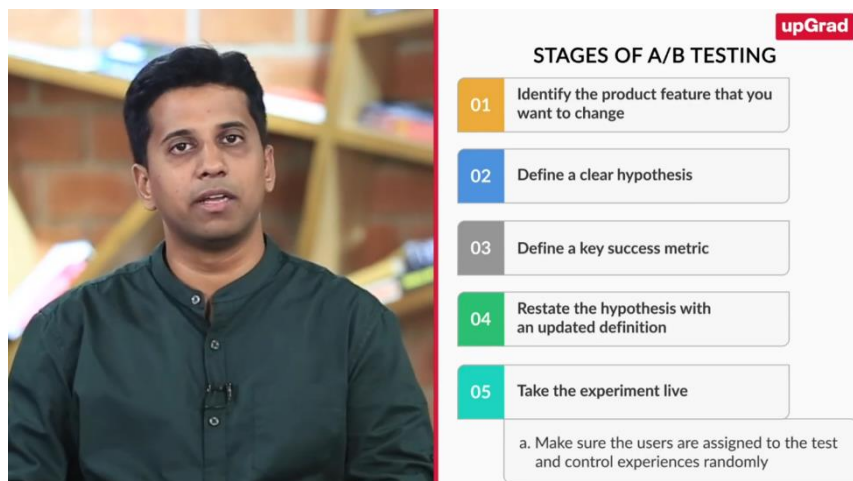
The second would be if the navigation is removed from the checkout pages, the conversion rates will increase because the analytics have shown that people drop off while navigating between pages.

Now out of these two, the second one is a stronger hypothesis because it is more detailed. Also has a rationale behind the test included. So, if the hypothesis is proved wrong, you know that the assumption you took based on the data was wrong.



So, now that we appreciate the importance of an AB testing and AB framework, let's understand the different stages of an AB experiment and how we would go about designing an experiment.

In the first part, let's look at the different stages in an AB experiment, what life cycle of it is. And in this section, there will be a couple of new terms which you haven't seen before. This is because of the depth we are going in. While it's fine, we'll explain these terms along the way.

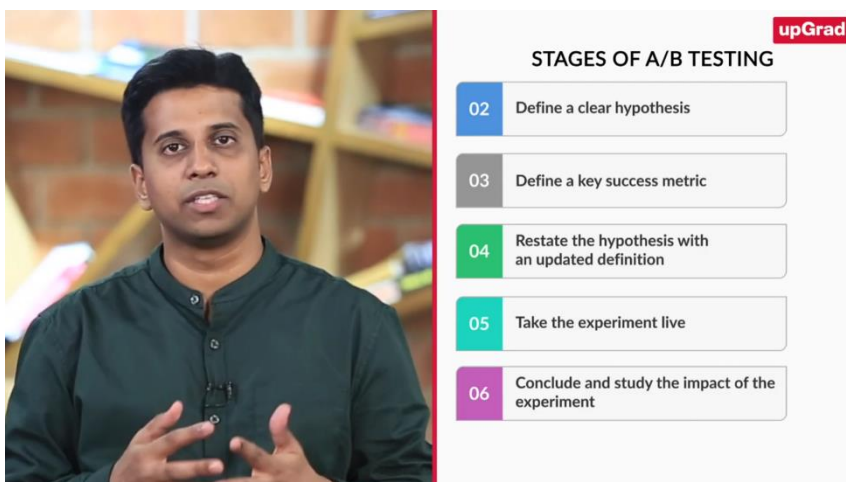


So, what are the different stages or the different big concentrations:

- First, the feature. What feature are you changing? What variation are you looking at? And we understand that you could be changing maybe just one feature or you could be changing combinations, or you could do any of

that. But ideally it is better that you work on one change. And the reason is that otherwise it is difficult to measure the impact of each change if you have multiple.

- Now that we've looked at the feature or finalized a feature you're changing, it's very important to make a clearly defined hypothesis. So, which should be a statement of the change, the kind of impact you see, and the rationale as to why do you expect that to work.
- The third part is where you define a key success metric. Now between second and the third, there could be a little bit of iteration because you have an objective in mind and you choose the metric and so on, but let's go with this flow for now that you have the hypothesis and you have the key success metric.
- And then once you have the key success metric, you just reiterate or restate the hypothesis with the updated definition. So, we have the key success metric, which comes in. And what is the key success metric? This is that one metric to help you assess whether the experiment worked or not, or that one metric on which you want the impact to occur.
- This metric, the key success metric is also important to help you identify what size of the sample you need to work at. And this therefore dictates the duration of the experiment as well. Also, it is best to have a bunch of supporting metrics.
- So, while your key success metric could be, say a click through rate, but there are some other guardrail metrics, which you don't want to affect. These good be around, let's say the overall conversion, or they could be on profitability and so on. But the best setup is when you have one key success metric and you have some guardrail metrics or supporting metrics to measure.
- Now we need to take the experiment live. And one of the key aspects here is making sure users are assigned to the test and the control experience randomly. This you want to ensure so that there is no bias in the test and the control. Essentially, should not be any other source of difference between the behaviours of these groups.

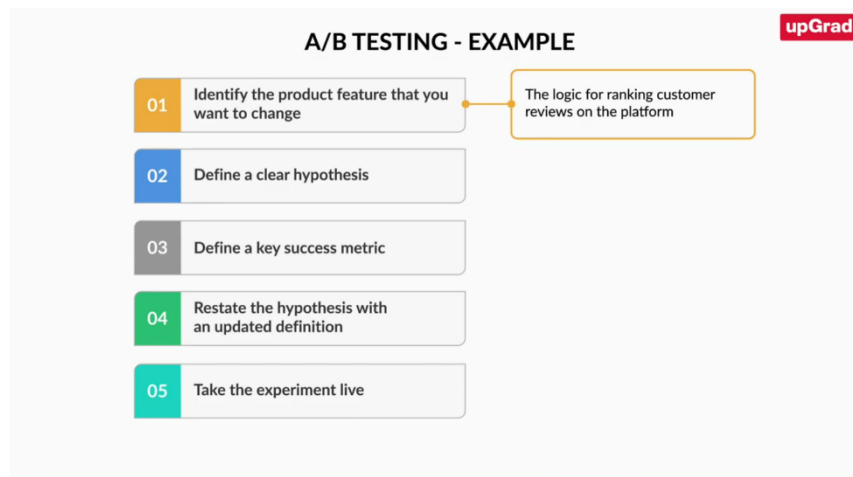


The graphic overlay, titled "STAGES OF A/B TESTING" with the upGrad logo, lists the following steps:

- 02 Define a clear hypothesis
- 03 Define a key success metric
- 04 Restate the hypothesis with an updated definition
- 05 Take the experiment live
- 06 Conclude and study the impact of the experiment

- Finally, after the entire experiment goes live, and after it has run for the stipulated, determined duration, you want to conclude the experiment. You want to study your impact on the key success metrics, how the supporting metrics moved and a post assessment to say whether this worked or not, where it worked, where it failed and so on.

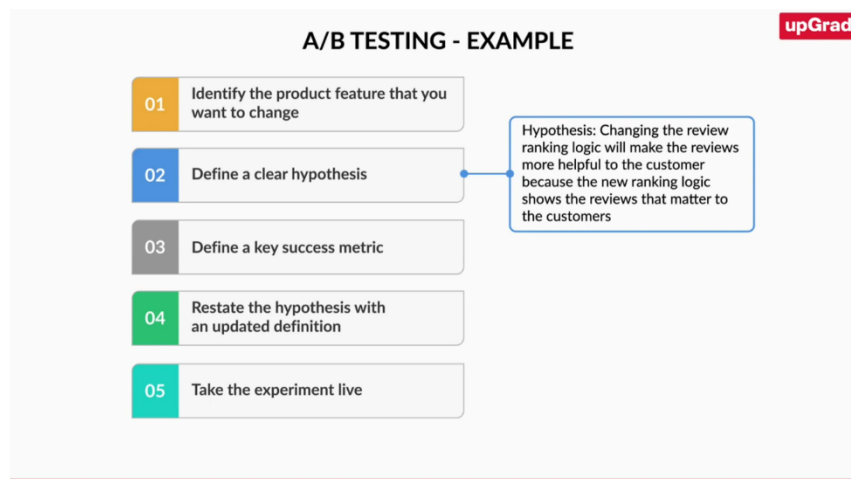
So, these are briefly the different stages in an AB experiment. We will look at each of these stage in detail with the example of the reviews module we saw earlier. Now let's take the reviews example and look at each step-in detail, how you would design the entire experiment.



So, the first page was the feature, identifying the feature. And in this case, let's say, we are changing the logic with which we rank the reviews. Essentially, when you go to any platform like Amazon or Flipkart, or even say a TripAdvisor and so on, you see reviews coming in for the item. In Amazon, it has reviews for that particular product.

Now there is some logic behind the scenes, which tells you that this review should come first, this review should come second and so on. Now you as a product manager, Amazon looking at the review's module have to decide on using or finding the best ranking algorithm for these reviews.

And you have some new algorithm which takes into consideration the customer's profile as well. So, essentially you can think of this as personalizing the review rankings for the customer. So, you have the existing state where things are not personalized, and you have a new contender or the new algorithm, which personalizes. So, the feature in this case is the review ranking logic.



So, now that we have the feature, what is our hypothesis? How do we state that? Broadly, the hypothesis at this stage can be stated as changing the ranking logic will make the reviews more helpful to the customers because the new ranking logic shows the reviews that matter to the customer.


So, note that the hypothesis still is not very specific or the way we want it to be. We have still not mentioned the key success metric, and we have not mentioned the expected impact. So, let's do that. Let's go into the discussion that own the key success metric and the expected impact. And then we will, again, restate the hypothesis based on what we arrive at as a key success metric and the impact.



The other aspect is what sort of impact are you looking, or what sort of change do you want to detect? This is called the minimum detectable effect. This is a new term. So, the minimum detectable effect or the MTE is by how much do you expect this change to impact your key success metric?

So, I mean, one good statement over here could be that if, let's say, I have a search click through rate that is my key success metric, minimum detectable effect would have to specify by how much I would want the click through rate to change.

So, you would say the key success metric could be click through rate. And the minimum detectable effect could be 2% or 200 basis points. But how do you arrive at this? So, this is a very tricky number to arrive at. And there is no rule to come to this. It's a lot of art business understanding hand science over here.



upGrad


PREREQUISITES FOR ESTIMATING MDE

- 01 Business understanding
- 02 Product knowledge
- 03 Perspective from past A/B experiments

And the business understanding is a very strong element here, how the product works. Maybe your perspective from past experiments or maybe it is just based out of actionability. For example, if the clickthrough increases from 37 to 37.1%, maybe you won't action on it. Maybe that change is not enough for you to invest your resources in implementing that change.

There are a lot of considerations around this, and based on all of these, you would come with one number saying that I would want to detect a change of say 20 basis points and so on.

So, we mentioned that based on the key success metric and the minimum detectable effect, you would arrive at the sample size and the duration. Let's talk about that now.



upGrad

UNDERSTANDING MDE

- 1 The amount of change that you want to bring about in your key success metric
- 2 If click-through rate (CTR) is the key success metric, then MDE is the change that you want to bring about in the CTR
- 3 The change in the key success metric during A/B testing should be statistically significant

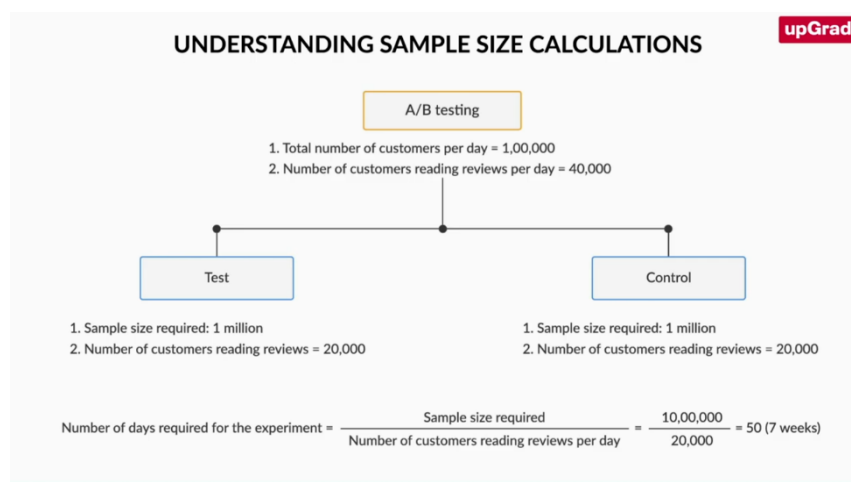
A key idea over here is that whatever change you see in the key success metric, that has to be statistically significant. Meaning that you should be able to be confident that change in the metric or this improvement is not purely by chance, that this is statistically sound and this kind of change will hold even after we move to this new change completely.

The detailed calculation of this will require heavy statistics. And it is better done by your data science or data analytics team.

But one thing to really understand over here is that statistical significance is not the same as business significance because a change of say 10 basis points on a click through rate, like from 37 to 37.1 may come as statistically significant, but maybe the business doesn't care about that minute change. So, for business, it may not be significant.

So, the duration of the experiment depends on the sample size you need. So, this has to be the relevant sample. What do we mean by that? Well, let's say you're doing an experiment around ranting in this case. And not all the customers, which come to the platform actually go and read reviews.

So, the relevant sample for you is the people who actually go and read the reviews. And that is to be factored in your calculation.



So, let's say, based on the current state of the key success metric and the change you want to detect, you get that you want 1 million as your sample size for test and 1 million for control. If this is given to you, you need to calculate the duration for the experiment. And you'd have to do some estimates over here.

So, let's say you have one lakh customers or a hundred thousand customers coming in daily onto the platform. And let's say about 40% of them would go ahead and read the reviews. Now that means 40K people every day, which are reading the reviews. And if you did test and control split, a 50, 50 split, 20 K every day will read the reviews in the control environment, 20K in the test.

And remember that you need 1 million samples for this experiment or to be conclusive. So, 1 million is what you want and you have 20K every day. So, you can just back calculate that, you need about 50 days as your duration for the experiment or about seven weeks.

So, that's how you calculate the experiment duration from the sample size. And depending on say, how deep down the funnel your metric is, you need to essentially back calculate all the way to the top of the funnel, depending on the traffic, what duration you need to run this experiment for.

Sample Size Calculator | Chi-Squared Test | Sequential Sampling | 2 Sample T-Test | Survival Times | Count Data

Need A/B sample sizes on your iPhone or iPad? Download [A/B Buddy](#) today.

Question: How many subjects are needed for an A/B test?

Baseline conversion rate: % 35% [\[link \]](#)

Minimum Detectable Effect: % 30% - 40%

The Minimum Detectable Effect is the smallest effect that will be detected 1-β% of the time.

☒ Absolute ☐ Relative Conversion rates in the gray area will not be distinguishable from the baseline.

Sample size:
1,440
per variation

Statistical power 1-β: % Percent of the time the minimum effect size will be detected, assuming it exists

Significance level α: % Percent of the time a difference will be detected, assuming one does NOT exist

See also: [How Not To Run an A/B Test](#)

So, with that, let's look at an example of how we do some basic calculations for getting to the sample sites. And for this purpose, we have many tools available and the calculations can get very involved as well, using statistics and so on.

But this is one tool which is very simple to use tool. And we can just enter the two values we discussed earlier, the key success metric and the minimum detectable effect. So, in this tool, the key success metric is over here, is mentioned as a baseline rate. So, this is some sort of an action rate, which is why it can be called some sort of a conversion rate over here.

So, all we need to do is enter the baseline conversion rate or the key success metric, the current state, and effect we want to capture or detect. And you may notice that there are a couple of sliders below. These are to do with the statistical part of it. So, we will not bother, and we will use the defaults.

We want to understand how the success metrics magnitude and the detectable effect affects the sample size. So, let's take a simple example where we have a click through rate of a company at about 35%. And we want a minimum detectable effect, or the change you want to detect is let's say 5%.

Sample Size Calculator | Chi-Squared Test | Sequential Sampling | 2 Sample T-Test | Survival Times | Count Data

Need A/B sample sizes on your iPhone or iPad? Download [A/B Buddy](#) today.

Question: How many subjects are needed for an A/B test?

Baseline conversion rate: % 35% [\[link \]](#)

Minimum Detectable Effect: % 34% - 36%

The Minimum Detectable Effect is the smallest effect that will be detected 1-β% of the time.

☒ Absolute ☐ Relative Conversion rates in the gray area will not be distinguishable from the baseline.

Sample size:
35,781
per variation

Statistical power 1-β: % Percent of the time the minimum effect size will be detected, assuming it exists

Significance level α: % Percent of the time a difference will be detected, assuming one does NOT exist

See also: [How Not To Run an A/B Test](#)

So, in this case, the sample size, which we need is 1,440 per variation. So, on 35% detecting, 5% absolute change needs just 1,440 samples per variation. Let's see the effect of changing the minimum detectable effect.

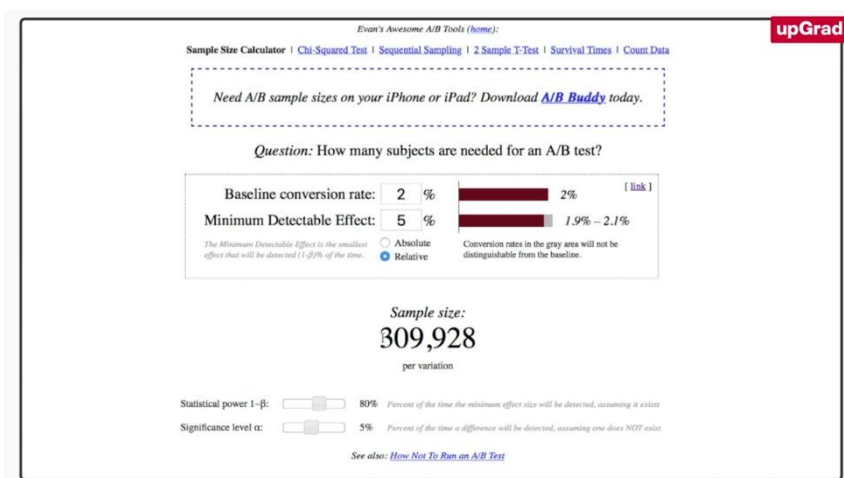
If I want to capture a smaller change, because 5% seems very ambitious, let's be more realistic. If I want to capture a smaller change, say 1%, then you see that the sample size requirement just explodes. It's from 1,440 to about 36,000. And if you want to detect an even smaller change, which is, let's say 50 basis points or 0.5% overall, you see that the sample size requirement is even higher. It is now close to one 43,000.

So, for the same baseline conversion rate or for the same key success metric, the minimum detectable effect has a big impact on the sample size required. The smaller the change you want to detect the larger, the sample size you need.

Now let's look at the other side of things. Let's try to see how the magnitude of the baseline conversion rate affects the sample size. So, what we'll do is we will fix a minimum detectable effect at say, 5% of the baseline or the key success metric. And we'll see how changing the magnitude of the key success metric changes the sample size requirement.

So, for now, we have 35% key success metric, and we have 5% relative minimum detectable effects, which is 5% of 35. And for this, the sample size requirement is about 11 point, 7,000.


Now, if I change the key success metrics magnitude, if let's say instead of having 35% CTR, let's say we were looking at some conversion metric could be, let's say at 4%, and you want to detect 5% relative change on that 4%. You see that the sample size has jumped from 11.7 to 152 approximately, 152,000 samples.



This is a big change. And if we further wanted to study something around, let's say a baseline conversion rate of 2%, and we want the relative 5% change, we see that the sample size needed is even more, which is about 310,000 per variation.

So, the second lesson is that the magnitude of your metric or the key success metric also has a big impact on the sample size requirement.

But you still haven't answered the question, what will be our key success metric for this review ranking case? So, let's discuss that. Let's first appreciate that one single feature change can actually impact multiple metrics. There's not just one single metric which this feature will back.



upGrad


CONSIDERATIONS FOR CHOOSING METRICS

- 1 Choose a metric that is directly impacted by the feature
- 2 While choosing between two metrics, choose the one with a lesser sample size requirement

So, you will often have a situation where you have multiple metrics to choose from.

- So, how do we choose in general terms? There are a couple of considerations which should help you choose the key metric. The metric, which is directly impacted by the feature or the metric which is closest to the feature. The most relevant metric is what you have to choose. That is the first consideration.
- The second is that in some cases, you may still be having a choice between a couple of metrics. In that case, the metric, which helps you get a significance or the results with a smaller sample size should be chosen. And this is of course, because if the metric takes a long duration, you may not have all that time, but we get the result and so on.

These are two considerations, but let's think in more concrete terms with our reviews, ranking logic example.




upGrad

WHY ARE REVIEWS CRUCIAL?

- 1 Build customer's trust
- 2 Improve repeat behaviour
- 3 Reducing returns

For our reviews example, reviews can help in many ways. Reviews can help build customer trust, reviews can help with repeat behaviour because they now value the platform, they visit the platform again and again, because they now feel that, you know, this platform is reliable for reviews and so on.

If the reviews are very helpful and reflective of the product, it should also help you reduce the returns. And also, people who find the reviews helpful will engage more with the review's ecosystem as well. So, there are so many things, right? Which metrics should you choose, the metric which is, or let's say the theme which is more relevant, more useful is around engagement?



upGrad

METRICS FOR CUSTOMER ENGAGEMENT

Metric 1 = $\frac{\text{Number of helpful clicks}}{\text{Total review views}}$

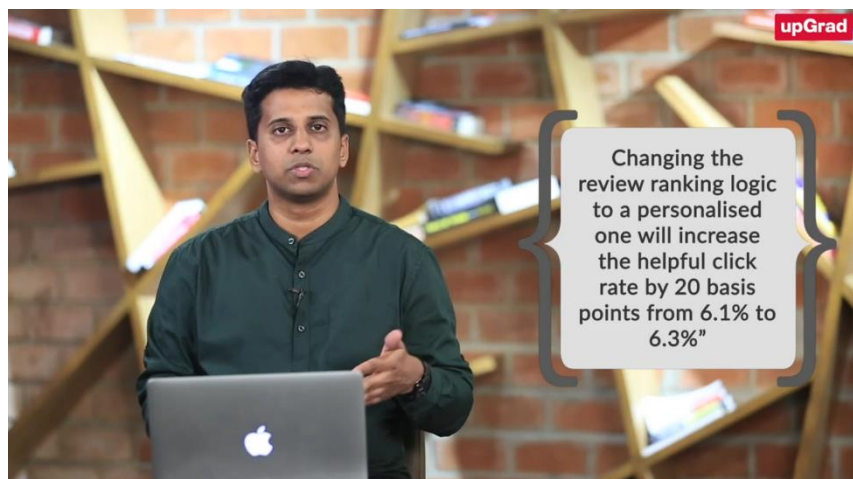
Metric 2 = $\frac{\text{Number of customers who found a review useful}}{\text{Number of customers who saw that review}}$

- People who found the reviews helpful or found the ranking logic better, they will click on the helpful button more often. That's what your hypothesis is. And you want to capture this in some metric. And there's a couple of metrics possible. You could say helpful clicks, divided by total review views, that is one metric which just tells you if all the reviews shown how many prompted a helpful click, that is an overall metric.
- Then you can also have a customer level metric, which something like, customers who found any review helpful divided by all customers who saw reviews. So, this is a measure of what proportion of customers are finding the reviews useful.

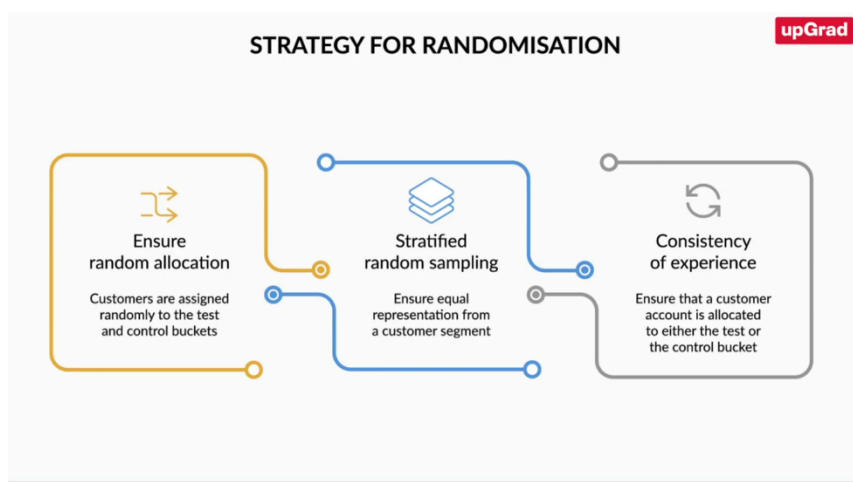
So, between these two, it's again a subjective call. It's a business call as well. You can finally take a call saying that, you know what, let's stick with the metric, which is more around the customers because the customers can see multiple reviews and so on.

The end objective is that the customer finds these reviews helpful. So, we will stick with the metric, which is customers, which found useful, divided by all customers who saw reviews. So, this could be our key success metric, and we also need to provide a minimum detectable effect.

And assuming that we did that exercise, let's say we arrived at some number around 20 basis points. So, that's a key success metric, which is customers clicking useful, divided by customers who saw a review. And the minimum detectable effect on this is 20 basis points.



So, with the key success metric and the minimum detectable effect identified, we now are able to restate the hypothesis in the format we need. The restated hypothesis now is changing the review ranking logic to a personalized one, will increase the helpful click rate by 20 basis points, from 6.1% to 6.3% because the reviews are more customized.



So, let's talk of the next step in this entire procedure, which is a randomization. And this is a very important step because if this is not done right, your entire analysis becomes void. Randomization should ensure that users are assigned to your test and the control groups completely randomly. There should be no bias or no systematic allocation to these buckets.

And in this randomization, there are a few strategies. One of them will be a new term, but I think the intuition is straightforward. You can have a simple, random sampling or a simple random allocation saying, you know what, every ordered customer, just through it to this bucket or the other.

It's a very simple randomization. Or you could have a stratified random sampling. What is stratified? Well, stratified is where you want to have equal representation from some user groups or segments. Essentially let's say in your experiment, you believe that Metro customers may behave or respond differently than non-Metro customers.

You don't want that to skew your analysis. So, what you do is you just make sure in the control, as well as the variants or test experiences, you have the same representation from the Metro, as well as non-Metro groups. So, this is what

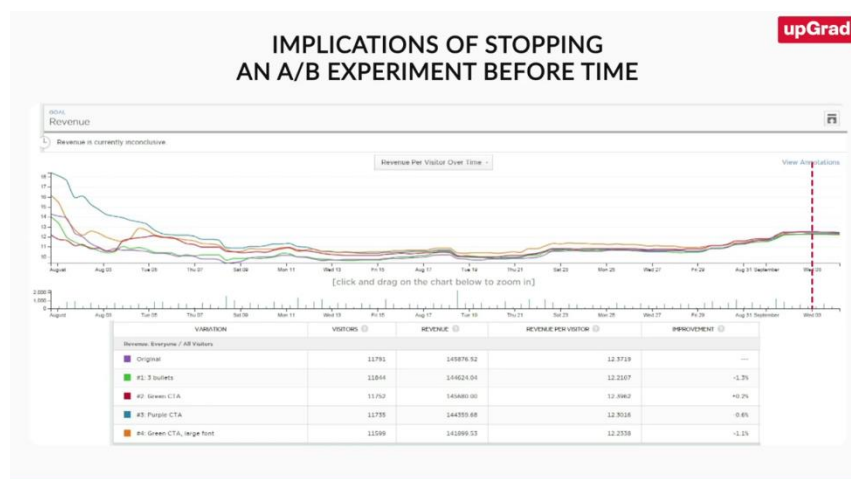
is called a stratified sampling, where you just ensure that you have a similar or equal representation of the groups in both the test and the control experiences.

And one more consideration in this randomization bit is consistency of experience for the customer. So, essentially randomization can happen at a visit level where say each visit of the customer is treated as a separate, let's say point, to allocate to a test or a control bucket.

So, what can happen is that I log into the platform now, I can see the test experience, and let's say, I log in again, after a few hours, I see either control experience.

Now this may not be ideal in many situations. So, it's always safer to randomise on the customer itself. So, maybe the customer or the account is allocated a bucket and the customer will always be allocated that group for the duration of the experiment.

So, if the experiment is running for a month, every time I come on to the app, I should see the same experience because otherwise it leads to differences in how I respond. And you cannot really analyse, well, what happened. Stopping before time can be very dangerous.



So, we have this view here where you had multiple say experiences being tested. And there was the original experience of course. And we can see in this graph with time, how the performance of each experience is. And when we began, when we began on day zero, I mean, right after day one, you can see that there is a big difference between the different variants.

You see that the blue one is it there, a top blue bar, which corresponds to the purple CTA. Is there a top, whereas your original call to action is somewhere in the middle? And the difference between them is huge. So, if on day five or day four also, if you had say assessed this or measured this, you would probably have concluded that blue is significantly above purple.

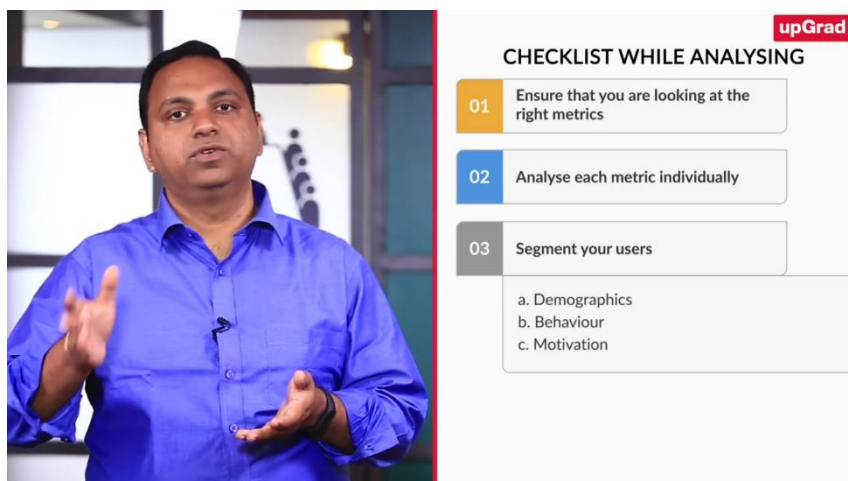
And even if we did a statistical test, you probably would have concluded it is statistically significant as well. And you could have stopped this experiment then with the conclusion that blue is the best.

But if you hadn't stopped and you continued a few more days, somewhere say around in this case the middle of it, you'd see that the original one is a lowest, purple line is the lowest and the highest line, the best line seems to be the saffron one.

And even now, the difference is significant. And enough days have passed. And now if you had stopped this, you would have concluded even based on statistical significance that the saffron line is the best. And again, your conclusion would have been off.

If only you continued towards the end, then you would have realized that all the experiences are pretty much the same in terms of your goal or the key success metric. And none of the variants has significantly outperformed your original experience.

So again, given enough time in the conclusion is very different, but if you had stopped at different points in the experiment, you would have concluded incorrectly and this topping, just because you see a big difference, and it is statistically significant, this is probably the most common error committed by people who are beginning. And this is one pitfall to avoid.



Once you have reached the required sample based on the data, you will have two options. If variation B wins, implement the change for all your users and launch the new version. If control A wins, identify what was wrong with the hypothesis and how it could have been improved.

While you analyse the data, you should always keep a few things in mind. Number one, check if you are looking at the right matrices. Also, if the test involved multiple matrices, you would need to look at each of them individually.

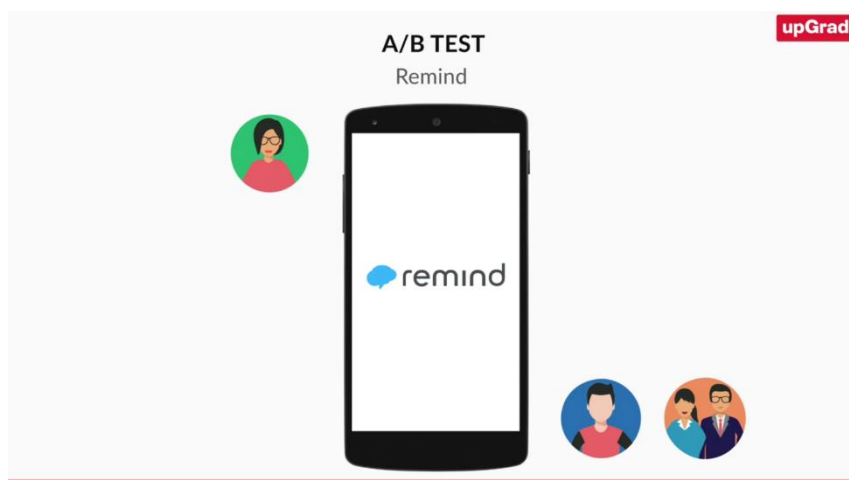
Segment your users, post the test based on common demographics or behaviour, you know, or motivation. In this way, you would be able to find out that the variations worked for the returning visitors, but not for the first-time visitors.

You should delve deeper in the user behaviour analysis. You could use techniques like heat map, click testing, visitor recording, to gather further insights on the AB test.



Recording your results properly is also important. One of the ways this can be done is by creating a detailed spreadsheet and sharing it with everyone else. You could include these details.

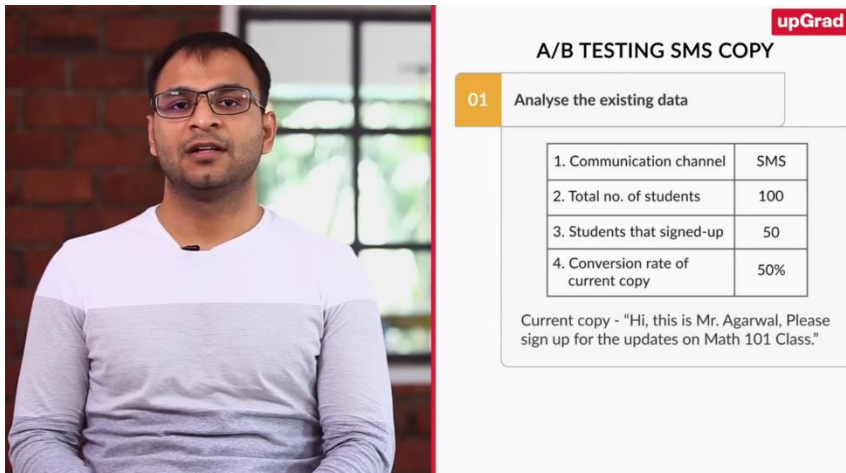
Start and end date, hypothesis, success matrix, what is your confidence level and what are the key takeaways of the AB test. And a link to the detailed report with the things like hypothesis, matrices tracked, details about the variation and the data of the experiment represented through charts and graphs.



Let's take an example of a real-life AB test that I conducted. So, a couple of years ago, I was working for a company called Remind. Now Remind is an app that helps teachers to communicate with students and parents.

So, basically the way the product works is we acquire a teacher and then we ask the teacher to send an invitation to all her students so that they can sign up and join her class. Once they join her class, teachers can send a broadcast message to all the students.

So, one of the key parts, as you can understand is the conversion rate from once a teacher is signed to when she gets her entire class signed up on her platform.



upGrad

A/B TESTING SMS COPY

01 Analyse the existing data

1. Communication channel	SMS
2. Total no. of students	100
3. Students that signed-up	50
4. Conversion rate of current copy	50%

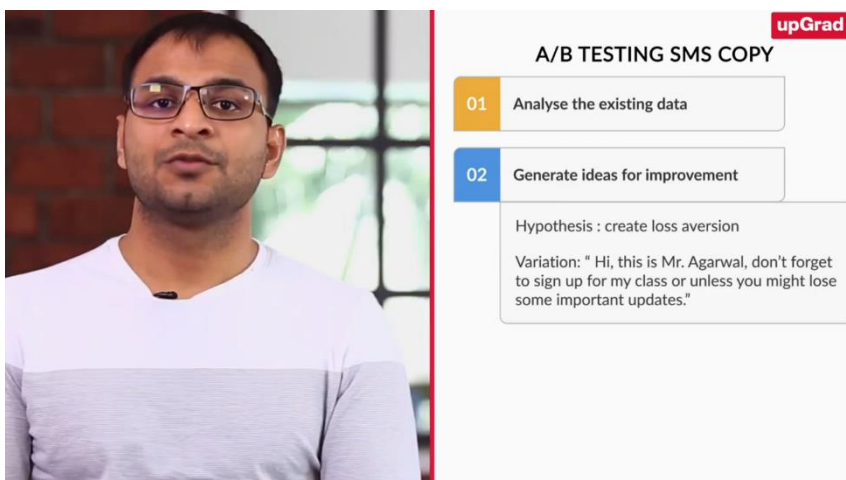
Current copy - "Hi, this is Mr. Agarwal, Please sign up for the updates on Math 101 Class."

The AB test that we decided to conduct was the SMS copy that we sent on the teacher's behalf as the invitation to the students. This test was very important because we knew that once a teacher gets a critical mass of students signed up for her class, then she became very actively engaged and she will never churn year on year.

So, it was very critical for us to help the teacher get all her students signed up for the class. And we were seeing that only 50% of the current copy of the SMS was converting. And SMS was the number one channel that the teacher used to spread the invitations.

So, when we looked at the copy of the SMS, it was a very straightforward SMS which said that, hi, this is Mr. Aggarwal, please sign up for updates for my math one on one class, and the link to join the class. So, we were thinking that, can we make this better? Because we knew that there's a lot of opportunity, there's only 50% conversion rate and there's a large volume of invitations.

So, that is the idea justifying that yes, there is an opportunity and that's the first step before doing any AB test. You have to first do a very simple math to assert that there is even an opportunity to put improvement.



upGrad

A/B TESTING SMS COPY

01 Analyse the existing data

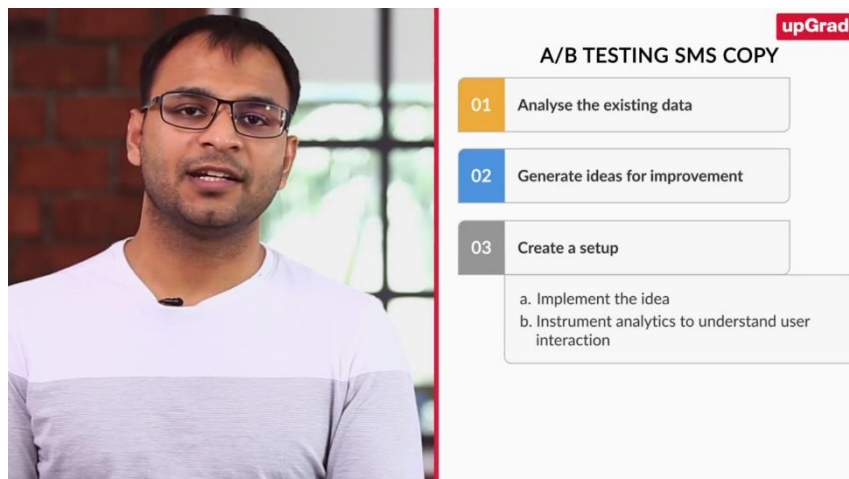
02 Generate ideas for improvement

Hypothesis : create loss aversion

Variation: " Hi, this is Mr. Agarwal, don't forget to sign up for my class or unless you might lose some important updates."

Then the second step was to generate ideas for improvement. One of the ideas we got was to create a loss aversion. It's very popular psychological term where instead of incentivizing people for something they can receive, if you tell people to avoid a punishment, they react to it more prominently.

So, one idea we generated for the copy was instead of saying, hi, this is Mr. Aggarwal, please join my class. We said hi, this is Mr. Aggarwal. Don't forget to sign up for my class or unless you might lose some important updates. So, this was the hypothesis for the test that if you do loss aversion, maybe the conversions rates increase.

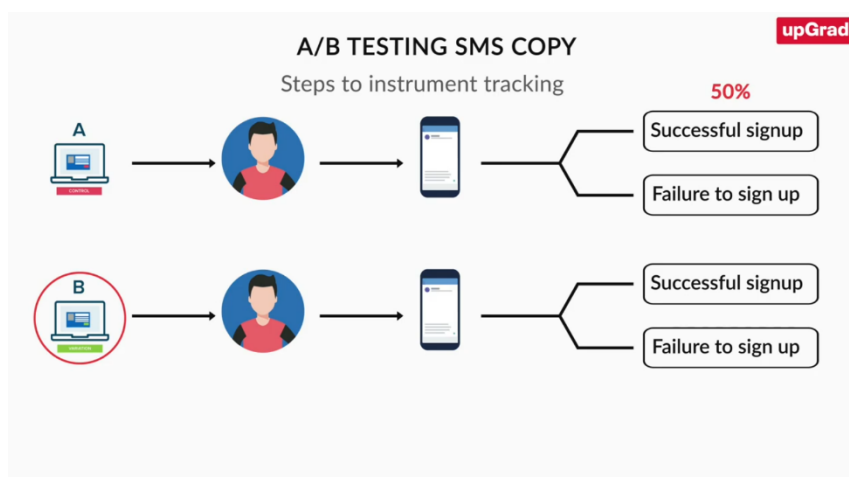


So, once we all agree that this is a good idea to test out, the next step was to create a setup. What is the set up supposed to do? The setup is supposed to do only two things.

A, it is supposed to implement your idea so that the users can actually interact with it. And the second thing is to instrument analytics so that you can understand how the users interacted and you can run, use the data to analyse whether the test was successful or not.

So, first step was implementation. For the implementation, what I did as a PM was, I wrote the two copies, the current copy and the text copy. The current copy is called the control, and the test copy is the loss aversion copy.

Once I wrote the two copies, I ask my engineers to decide to split the traffic of all invitations, 50-50 and then I also included instructions on how to instrument tracking of each step in the conversion rate. And the last step was to instrument the tracking for each step in the conversion funnel.



So, what loosely the funnel looks like here. The first step is teacher sends invites, say the teacher sends a hundred invites. The second steps were invites were sent from our system, just to make sure that there were no technical glitches. The third step were invites were delivered.

The fourth step is that students clicked on the link to sign up for the class and the fifth step is students successfully signed up for the class.

So, these are the five steps that we needed tracking for, and we implemented the tracking for control and test. So, basically, I could see the number of invitations that were sent to the control group, the number of invitations that were sent to the test group, and all subsequent metrics segmented by the control in the test group.

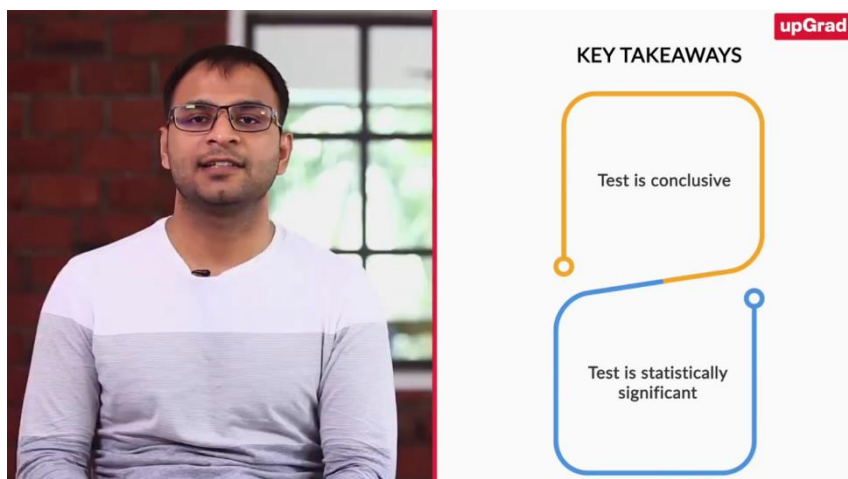
Once these metrics were set up, then the engineering just ran, taken the test life. What that basically means is that now, for all current invitations that are being sent, half of them will be send the control copy and half of them will be sent the test copy. And we will get real time analytics on the number of invites been sent and the number of invites been accepted.

Next, it was time to look at the results and the learnings from the test. So, the results were very clearly in the favour of the test group. We saw that the conversion, the base control rate was still at 50% as in 50% of people who received the first copy accepted the invite.

But for the loss aversion thing, it had jumped to 77%. That meant that people responded a lot more favourably to the loss aversion copy, hence confirming the very popular psychology experiment.

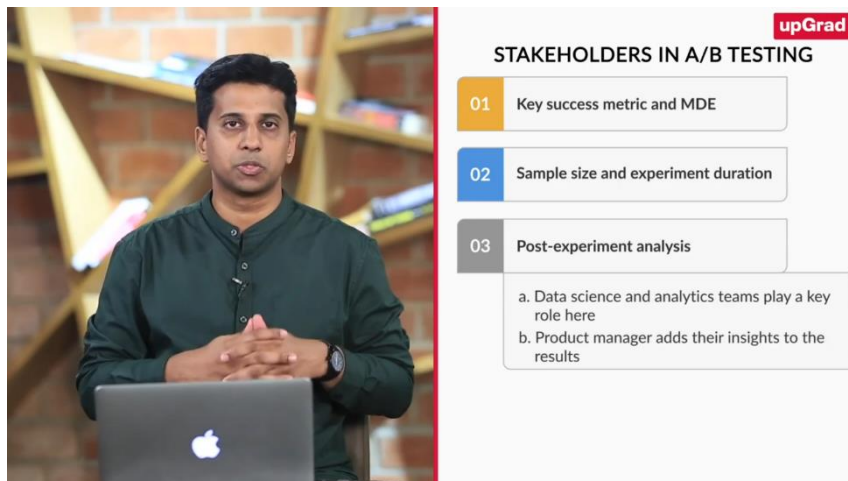
Our learning from the test was that A, that we should replace all the control copy with the test copy, and B, since we were able to get the lift from 50 to 77%, there is probably a lot more here we could do. Maybe we could generate more ideas to create even meaningful copy. Like what if we include the name of the student in the invite.

So, instead of saying, hi, this is Mr Aggarwal, please join my class. We could say, Hi Prateek, this is Mr. Aggarwal, please join my class. Because people would respond to the personal message more. And hence, we decided to generate a lot more copies to get the conversion rate even higher from 77%.



So, while this test was successful, there are quite a lot of things you need to keep in mind while running an AB test. The most important thing that you have to keep in mind is the test is, A, conclusive, and B statistically significant.

What that means? That means that you have tests, you run the test long enough and exposed enough people to the control and test group that whatever benefit you've got in the end is actually material, statistically significant and will hold if you move all the people to the winning variation.



So, we saw all the different steps in an AB experiment, the different stages of it. And one question could come is that what parts does the product manager own? And, you know, not all product managers are conversant with statistics, and so on, which part should they play an active role, and because product managers will also have some analytics or data science team supporting them, which parts can they play.

So, let's talk about that. Let's talk about the major considerations in an AB experiment and which stakeholder plays what role. While essentially, where does a PM play the primary role versus where does the PM play the secondary owner.

1. So, the first part is the key success metric and the minimum detectable effect. This really comes from understanding of the domain, understanding of the product, and so on. Also, depends on the business actionability.

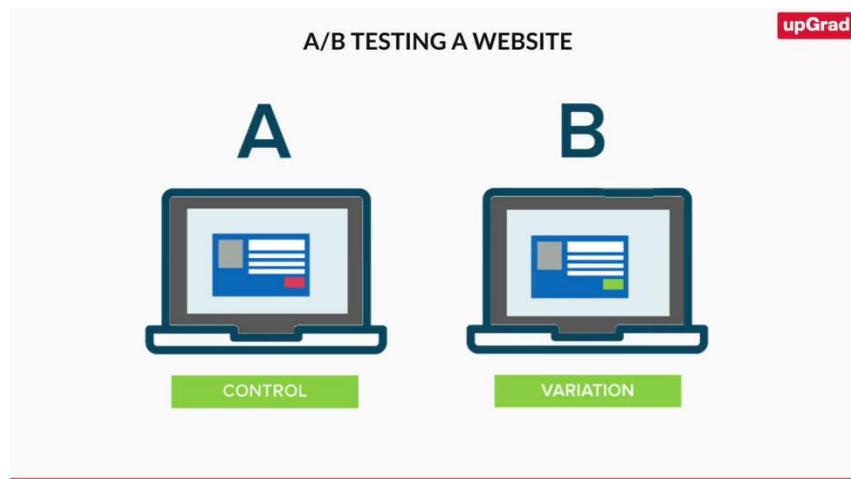
So, this is the part where the product manager plays the key role. You are the primary person in this decisioning. Of course, you will need some data points. You will need some analysis from the data science or analytics team to help you arrive at these numbers.

2. The second major consideration is the sample size and duration. Once the KSM and the MD or the key success metric, and the minimum detectable effect have been identified, now comes the statistics part where the sample size has to be estimated. And that affects the duration of the experiment.

Now, in this part the data science or analytics team will play the major role because these calculations can get rather involved. So, what they would do is given these two inputs, they will calculate how many weeks or what number of days you need to run this experiment for.

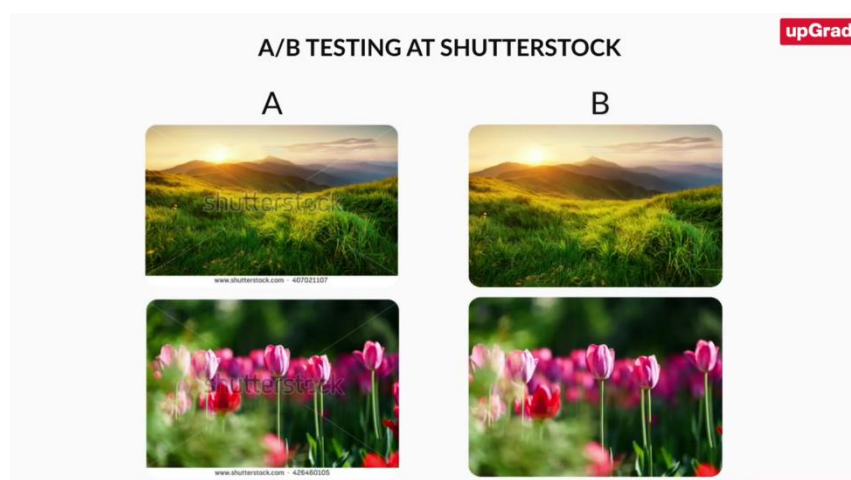
And once the experiment has been launched and all the tracking has been set up, we need to validate whether it's working fine and numbers are being tracked correctly or not.

3. And after that, after the entire duration of the experiment has passed, we need to do a post analysis. We need to report the results of the experiment. And in this, again, data science or analytics will play a key role. But the product manager, or you will also be a co-owner in this, supporting this analysis. In your capacity, bringing in your insights, bringing in a better hypothesis to assess this entire experiment results.



Let us recall what we learned in this session. We started the session with AB testing. Now, what is AB testing? AB testing is a method that allows you to test two different versions of the same element. The existing version is called control and the new design is called variation.

Traffic is divided between control and variation. After that, the performance of each design is measured with respect to the metrics that are important from a business point of view.

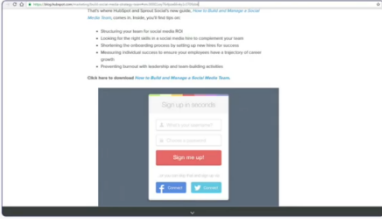


You learned from the example of Shutterstock, how they tested the hypothesis of disabling watermarks, and images.

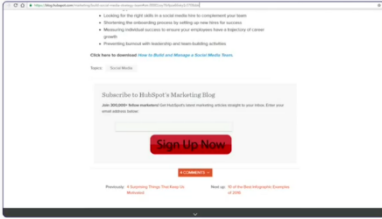
A/B TESTING AT HUBSPOT

Sign-up form test

71% increased Signups




A



B

We also saw another example where HubSpot, an inbound marketing company, AB tested two variations of presenting sign up forms to the users. Further, we understood how AB testing helps improve conversions. The test solely relies on data to generate an outcome, thus ensuring the credibility of the test.

After this, we understood the different elements that can be AB tested. These include text, layout, navigation flow, and even promotions and pricing tiers. So, these are the elements that can be tested.



CHECKLIST WHILE ANALYSING

- 01** Ensure that you are looking at the right metrics
- 02** Analyse each metric individually
- 03** Segment your users
 - a. Demographics
 - b. Behaviour
 - c. Motivation

Now, you also saw how to go about framing a hypothesis to test the elements. Once the test is run successfully, we should keep two things in mind when analysing the data. The first is whether we are looking into the correct metrics. And secondly, if there are more than one metric being tracked, we need to look at them individually.

Next to further refine the data, we learned that we should segment the users and the respective data. You can also use certain techniques like heat mapping or click testing to get as much data as possible.



Once you have the data, it is recommended to create a spreadsheet, outlining the various details, like the start and end date, the hypothesis, success metrics, and key takeaways of AB testing to name a few. So, that is all for this session. See you in the next one.

No part of this publication may be reproduced, transmitted, or stored in a retrieval system, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher.