# Nine Common A/B Testing Pitfalls and How to Avoid Them 

Jonas Dahl and Doug Mumford

Table of contents
1: Ignoring the effects of the significance level
2: Declaring winners of multiple offer tests with no statistically significant difference
3: Ignoring the effects of statistical power
4: Using one-tailed tests
4: Monitoring tests
5: Stopping tests prematurely
5: Not considering novelty effects
6: Not considering differences in the consideration period
6: Using metrics that do not reflect business objectives
7: About the authors

A/B testing forms the backbone of most digital marketing optimization programs, helping marketers offer optimized and targeted experiences to their visitors and customers. This paper outlines nine of the most significant pitfalls that companies fall prey to when performing $A / B$ testing. It also includes ways to avoid them, so your company can achieve greater ROI through its testing efforts and have greater confidence in its reported A/B test results.

## 1. Ignoring the effects of the significance level

How likely is it that your test reports a significant difference in conversion rate between two offers when in fact there is none? That's what the significance level of a test helps determine. Such misleading findings are often called a false positive, and in the world of statistics, are called a Type I error.

When you specify the significance level of an $A / B$ test, you're making a trade-off between your tolerance for false positives versus the number of visitors that the test must include.

Another directly related term, the confidence level, takes more of a glass half-full approach. Rather than stating the likelihood that you'll get a false positive, as the significance level does, the confidence level represents the likelihood that your test will not make that mistake.

Confidence level and significance level are directly related because:

$$
100 \% \text { - confidence level = significance level }
$$

In A/B testing, marketers often use $95 \%$ confidence levels. Clearly, based on the above equation, that corresponds to a significance level of $5 \%$. Testing with a $95 \%$ confidence level means you have a $5 \%$ chance of detecting a statistically significant lift, even when in reality, there's no difference between the offers.

As the graph below illustrates, the more tests that you run, the more likely at least one of those tests will result in a false positive. For example, if you run 10 tests using a $95 \%$ confidence level, there is approximately a $40 \%$ chance that you will detect one or more false positives. ${ }^{1}$


[^0]In a marketing organization, $95 \%$ usually constitutes a reasonable trade-off between the risk of a false positive and the opportunity costs of running the test for a longer time to have greater confidence in the results. However, two situations warrant paying close attention to the significance level and its implications for test results: post-test segmentation and testing multiple offers.

## Post-test segmentation

Marketers often slice and dice the results of a test based on visitor segments after the $A / B$ test concludes. Common segments include browser type, device type, geographic areas, time of day, and new versus returning visitors. This practice, known as post-test segmentation, provides excellent insight into visitor segments. In turn, marketers can use these insights to create better-targeted and more-relevant and differentiated content.

If there is no real difference in conversion rate, each time you test a segment, the probability of a false positive equals the significance level. And, as mentioned, the more tests you run, the greater the likelihood that you'll experience at least one false positive among those tests. In essence, each post-test segment represents a separate test. With a significance level of $5 \%$, on average you'll fall prey to one false positive every time you look at 20 post-test segments. The chart above shows how that likelihood increases.

Should you simply not do post-test segmentation? No-post-test segments are valuable. Instead, to avoid this cumulative false positive issue with post-test segmentation, after you've identified a post-test segment, consider testing it in a new test. Alternatively, you can apply the Bonferroni correction, which we discuss next.

## Testing multiple offers

Marketers frequently test more than two offers (or experiences) against each other. That's why you sometimes see $A / B$ testing solutions called $A / B / n$ testing, where $n$ is the number of offers that you are testing simultaneously.

It's important to note that each offer tested has a false positive rate equal to the significance level, as described above. Again, you are effectively running multiple tests when several offers are pitted against each other within a single test environment. For example, if you compare five offers in an $A / B / C / D / E$ test, effectively you form four comparisons: control to B, control to C, control to D, control to E. With a confidence level of 95\%, rather than the $5 \%$ probability of a false positive, you actually have $18.5 \%$. $^{2}$

To keep your overall confidence level at 95\% and avoid this issue (and impress people at a cocktail party), you apply what is known as the Bonferroni correction. Using this correction, you simply divide the significance level by the number of comparisons to come up with the significance level that you need to achieve a $95 \%$ confidence level.

Applying the Bonferroni correction to the example above, you would use a $5 \% / 4=1.25 \%$ significance level, which is the same as a $98.75 \%$ confidence level for an individual test ( $100 \%-1.25 \%=98.75 \%$ ). This adjustment maintains the effective confidence level at $95 \%$ when you have four tests, as in the described example.

## 2. Declaring winners of multiple offer tests with no statistically significant difference

With multiple offer testing, marketers often declare the offer with the highest lift as the test winner, even though there is no statistically significant difference between the winner and the runner-up. This situation occurs when the difference between the alternatives is smaller than the difference between the alternatives and the control. The figure below illustrates this concept, with the black error bars representing $95 \%$ lift confidence intervals. The true lift for each offer relative to the control offer is $95 \%$ likely to be included within the confidence interval-the range shown by the error bars.


[^1]Offers $A$ and $B$ have the highest observed lift during the test, and it would be unlikely that offer $C$ would outperform those offers in a future test, because the confidence interval of $C$ does not even overlap with the confidence intervals of $A$ or $B$. However, even though offer $A$ has the highest observed lift during the test, it is quite possible that offer $B$ could perform better in a future test because the confidence intervals overlap.

The takeaway here is that both offers $A$ and $B$ should be considered winners of the test.
It's typically not feasible to run the test long enough to identify the true relative performance of the alternatives, and oftentimes the difference in performance between the alternatives is too small to substantially impact the conversion rate. In such cases, you can interpret the outcome as a tie and use other considerations, such as strategy or alignment with other elements of the page, to determine which offer to implement. With multiple tests, you must be open to more than one winner, which in some cases considerably opens up the possibilities for the direction to take with your website development.

Note that if you do want to identify the offer with the highest conversion rate, you are comparing all offers to every other offer. In the example above, you have $n=5$ offers-you have to make $n(n-1) / 2$ comparisons, or $5^{*}(5-1) / 2=10$ comparisons. In this case, the Bonferroni correction requires that the significance level of the test be $5 \% / 10=0.5 \%$, which corresponds to a confidence level of $99.5 \%$. However, such a high confidence level might require you to run the test for an unreasonably long period of time.

## 3. Ignoring the effects of statistical power

Statistical power is the probability that a test will detect a real difference in conversion rate between offers. Because of the random, or as statisticians like to call it, "stochastic," nature of conversion events, a test might not show a statistically significant difference, even when a real difference exists in conversion rate between two offers in the long run. Call it bad luck or by chance. Failing to detect a true difference in conversion rate is called a false negative or a Type II error.

The single most important factor that determines the power of a given test is the sample size, that is, the number of visitors included in the test. The statistical power also depends on the magnitude of the difference in conversion rate you want the test to detect. Perhaps this is somewhat intuitive, but if you are interested in only detecting large conversion rate differences, there's a much higher probability that the test will actually detect such large differences-kind of like discovering that you have an elephant in your living room versus a fly while looking through a paper towel tube. Along those lines, the smaller the difference you want to detect, the more observations, and therefore, time to get those observations, you require.

Today's marketers underpower a remarkable number of tests. In other words, they use a sample size that is too small. That means that they have a slim chance of detecting true positives, even when a substantial difference in conversion rate actually exists. In fact, if you continually run underpowered tests, the number of false positives can be comparable to, or even dominate, the number of true positives. This often leads to implementing neutral changes to a site (a waste of time) or changes that actually reduce conversion rates.


To avoid underpowering your test, consider that a typical standard for a well-powered test includes a confidence level of $95 \%$ and a statistical power of $80 \%$. Such a test offers a $95 \%$ probability that you'll avoid a false positive and an 80\% probability that you'll avoid a false negative.

If you want more detail on this important subject, we'll soon be posting a paper entitled, Designing a WellPowered $A / B$ Test, that describes how to determine an adequate sample size based on these suggested requirements for confidence level and statistical power.

## 4. Using one-tailed tests

One-tailed tests require a smaller observed difference in conversion rates between the offers to call a winner at a certain significance level. This seems appealing because winners can be called earlier and more often than when using two-tailed tests. But in keeping with the saying, "There's no free lunch," one-tailed tests come at an expense.

In a one-tailed test, you test whether offer B is better than offer $A$. The direction of the test has to be determined before the test commences, or "a priori" in statistics-speak. In other words, you must decide whether to test for $B$ being better than $A$ or $A$ being better than $B$ before initiating the test. However, if you look at the results of the $A / B$ test and see that $B$ is doing better than $A$ and then decide to do a one-tailed test to see whether that difference is statistically significant, you are violating the assumptions behind the statistical test. Violating the assumptions of the test means that your confidence intervals are unreliable and the test has a higher false positive rate than you would expect.

You might view a one-tailed test as putting an offer on trial with a judge who has already made up his or her mind. In a one-tailed test, you've already decided what the winning offer will be and are out to prove it, rather than giving each experience an equal chance to prove itself as the winner. One-tailed tests should only be used in the rare situations in which you are only interested in whether one offer is better than the other and not the other way around. To avoid the issue of the one-tailed test, use an $A / B$ testing solution that always uses two-tailed tests, like Adobe Target.

## 5. Monitoring tests

Marketers frequently monitor $A / B$ tests until the test determines a significant result. After all, why test after you've achieved statistical significance?

Unfortunately, it's not that simple. Not to throw a wrench in the works, but it turns out that monitoring the results adversely impacts the effective statistical significance of the test. It actually greatly increases the likelihood of false positives and makes your confidence intervals untrustworthy.

This might seem confusing. It sounds like we are saying that just from looking at your results mid-test, you can cause them to lose their statistical significance. That's not exactly what's going on. The following example explains why.

Let's say you simulate 10,000 conversion events of two offers, with both offers having $10 \%$ conversion rates. Because the conversion rates are the same, you should detect no difference in conversion lift when you test the two offers against each other. Using a $95 \%$ confidence interval, the test results in the expected $5 \%$ false positive rate when it is evaluated after collecting all 10,000 observations. So if we run 100 of these tests, on average we will get five false positives (in actuality, all positives are false in this example because there is no difference in conversion rate between the two offers). However, if we evaluate the test 10 times during the test—every 1,000 observations-it turns out that the false positive rate jumps up to $16 \%$. Monitoring the test has more than tripled the risk of false positives! How can this be?

To understand why this occurs, you have to consider the different actions taken when a significant result is detected and when it is not detected. When a statistically significant result is detected, the test is stopped and a winner is declared. However, if the result is not statistically significant, we allow the test to continue. This situation strongly favors the positive outcome, and hence, distorts the effective significance level of the test.

To avoid this problem, you must determine an adequate sample size-the number of visitors exposed to each offer-before initiating the test. While it's fine to look at the test results during the test to make sure that you implemented the test correctly, do not draw conclusions or stop the test before the required number of visitors is reached.

## 6. Stopping tests prematurely

It is tempting to stop a test if one of the offers performs much better or worse than the others in the first few days of the test. However, when the number of observations is low, there is a high likelihood that a positive or negative lift will be observed just by chance because the conversion rate is averaged over a low number of visitors. As the test collects more data points, the conversion rates converge toward their true long-term values.

The figure below shows five offers that have the same long-term conversion rate. Offer B had a poor conversion rate for the first 2,000 visitors, and it takes a long time before the estimated conversion rate returns to the true long-term rate.


Number of visitors

This phenomenon is known as "regression to the mean," and can lead to disappointment when an offer that performed well during the initial days of a test fails to keep up this level of performance in the long run. It can also lead to lost revenue when a good offer is not implemented because it happened to underperform in the early days of a test just by chance.

Much like the pitfall of monitoring your test, the best way to avoid these issues is to determine an adequate number of visitors before running the test and then let the test run until this number of visitors has been exposed to the offers.

## 7. Not considering novelty effects

Other unexpected things can happen if we don't allow enough time for running a test. This time the problem is not a statistics problem; it's simply a reaction to change by the visitors. If you change a well-established part of your website, returning visitors might at first engage less fully with the new offer because of changes to their usual workflow. This can temporarily cause a superior new offer to perform less optimally until returning visitors become accustomed to it-a small price to pay given the long-term gains that the superior offer will deliver.

To determine if the new offer underperforms because of a novelty effect or because it's truly inferior, you can segment your visitors into new and returning visitors and compare the conversion rates. If it's just the novelty effect, the new offer will win with new visitors. Eventually, as returning visitors get accustomed to the new changes, the offer will win with them, too.

The novelty effect can also work in reverse. Visitors often react positively to a change just because it introduces something new. After a while, as the new content becomes stale or less exciting to the visitor, the conversion rate drops. This effect is harder to identify, but carefully monitoring changes in the conversion rate is key to detecting this.

## 8. Not considering differences in the consideration period

The consideration period is the time period from when the $A / B$ testing solution presents an offer to a visitor to when the visitor converts. This can be important with offers that affect the consideration period substantiallyfor example, an offer that implies a deadline, such as "Time-limited offer-Purchase by this Sunday."

Such offers nudge visitors to convert sooner and will be favored if the test is stopped immediately after the offer expires, because the alternative offer might have a longer deadline or no deadline, and therefore, a longer consideration period. The alternative would get conversions in the period after the termination of the test, but if you stop the test at the end of the deadline, further conversions do not get counted toward the test conversion rate.

The figure below shows two offers that two different visitors see at the same time on a Sunday afternoon. The consideration period for offer A is short, and the visitor converts later that day. However, offer B has a longer consideration period, and the visitor who saw offer $B$ thinks about the offer for a while and ends up converting Monday morning. If you stop the test Sunday night, the conversion associated with offer A is counted toward offer A's conversion metric, whereas the conversion associated with offer B is not counted toward offer B's conversion metric. This puts offer B at a significant disadvantage.


To avoid this pitfall, allow some time for visitors who were exposed to the test offers to convert after a new entry to the test has been stopped. This step gives you a fair comparison of the offers.

## 9. Using metrics that do not reflect business objectives

Marketers might be tempted to use high-traffic and low-variance conversion metrics in the upper funnel, such as click-through rate (CTR), to reach an adequate number of test conversions faster. However, carefully consider whether CTR is an adequate proxy for the business goal that you want to attain. Offers with higher CTRs can easily lead to lower revenue. This can happen when offers attract visitors with a lower propensity to buy, or when the offer itself-for example, a discount offer-simply leads to lower revenue.


Consider the skiing offer below. It generates a much higher CTR than the cycling offer, but because visitors spend much more money on average when they follow the cycling offer, the expected revenue of putting the cycling offer in front of a given visitor is higher. Therefore, an A/B test with CTR as the metric would pick an offer that does not maximize revenue-the fundamental business objective.


Expected Revenue = \$10


Expected Revenue = \$25

To avoid this issue, monitor your business metrics carefully to identify the business impact of the offers, or better yet, use a metric that is closer to your business goal, if possible.

## Success with A/B testing by recognizing and stepping around the pitfalls

After learning about the common A/B testing pitfalls, we hope you can identify when and where you might have fallen prey to them. We also hope we've armed you with a better understanding of some of the statistics and probability concepts involved in $A / B$ testing that often feel like the domain of people with math degrees.

The steps below help you avoid these pitfalls and focus on achieving better results from your $A / B$ testing:

- Carefully consider the right metric for the test based on relevant business goals.
- Decide on a confidence level before the test starts, and adhere to this threshold when evaluating the results after the test ends.
- Calculate the sample size (number of visitors) before the test is started.
- Wait for the calculated sample size to be reached before stopping the test.
- Adjust the confidence level when doing post-test segmentation or evaluating more than one alternative-for example, by using the Bonferroni correction.

For more information about using Adobe Target for optimizing and personalizing your websites and mobile sites, visit www.adobe.com/target.

## About the authors

Jonas Dahl, a business analyst with Adobe Target, is interested in the intersection between traditional and data-driven marketing. He is passionate about understanding customer needs and behavior through data and is engaged in developing best practices for testing and targeting. Dahl earned his MBA from the Yale School of Management, where he focused on behavioral economics and data-driven marketing. He holds a PhD in engineering from Aalborg University in Denmark.

Doug Mumford is an optimization consultant with Adobe and an expert on $A / B$ and multivariate testing with Adobe Target. He helps enterprise clients create and execute the strategies that improve the marketing effectiveness of their digital properties. As a thought leader in optimization, he co-developed the industry's most comprehensive optimization maturity model. Mumford holds both a BS in computer engineering and an MBA with a focus on marketing strategy from Brigham Young University.

[^2]
[^0]:    ${ }^{1}$ Given that there is no real lift: $\operatorname{Pr}($ at least one false positive $)=1-\operatorname{Pr}($ no false positives $)=1-0.95^{\wedge} 10=40 \%$.

[^1]:    ${ }^{2}$ Given that there is no real lift: $\operatorname{Pr}($ at least one false positive $)=1-\operatorname{Pr}($ no false positives $)=1-0.954=18.5 \%$.

[^2]:    Adobe and the Adobe logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. All other trademarks are the property of their respective owners.
    © 2014 Adobe Systems Incorporated. All rights reserved. Printed in the USA.
    $12 / 14$

