Testing is no joke – you have to test right. Bad testing is even worse than no testing at all – the reason is that you might be confident that solutions A, B and C work well while in reality they hurt your business.

Poor A/B testing methodologies are costing online retailers up to $13bn a year in lost revenue, according to research from Qubit. Don't take this lightly!

Very typical story of a business that does a/b testing is that they run 100 tests over the year, yet a year later their conversion rate is where it was when they began. Why? Because they did it wrong. Massive waste of time, money and human potential.

Here are key things to keep in mind for running tests:

Before you end the test you need 3 things: big enough sample size, long enough test duration (minimum 2 business cycles) and only then statistical significance

In order to be confident that the results of your test are actually valid, you need to know how big of a sample size you need. This is something you calculate up front.

There are several calculators out there for this – like this one.

You need a minimum number of observations for the right statistical power. Using the number you get from the sample size calculators as a ballpark is perfectly valid, but the test may not be as powerful as you had originally planned. The only real danger is in stopping the test early after looking at preliminary results. There's no penalty to have a larger sample size (only takes more time).

As a very rough ballpark I typically recommend ignoring your test results until you have at least 350 conversions per variation (or more – depending on the needed sample size). I repeat: **this is just a ballpark figure**. Calculating needed sample sizes is science, not magic! There are no magic numbers.

If you want to analyze your test results across segments, you need even more conversions (each segment needs to have valid sample sizes before you look at it). It's a good idea to run tests targeting a specific segment, e.g. you have separate tests for desktop, tablets and mobile.

Once your test has enough sample size and the test has run long enough (2-4 weeks), we want to see if one or more variations is better than Control. For this we look at statistical significance.

Don't stop the test just when you reach 95% confidence (or higher)

This is the first rule, and very important. It's human to scream "yeah!" and want to stop the test, and roll the treatment out live. Many who do discover later (if they bother to check) that even though their test got like +20% uplift, it didn't have any impact on the business. Because there was no actual lift – it was imaginary.

Consider this: One thousand A/A tests (two identical pages tested against each other) were run.

- ➔ 771 experiments out of 1.000 reached 90% significance at some point
- ➔ 531 experiments out of 1.000 reached 95% significance at some point

Quote from the experimenter:

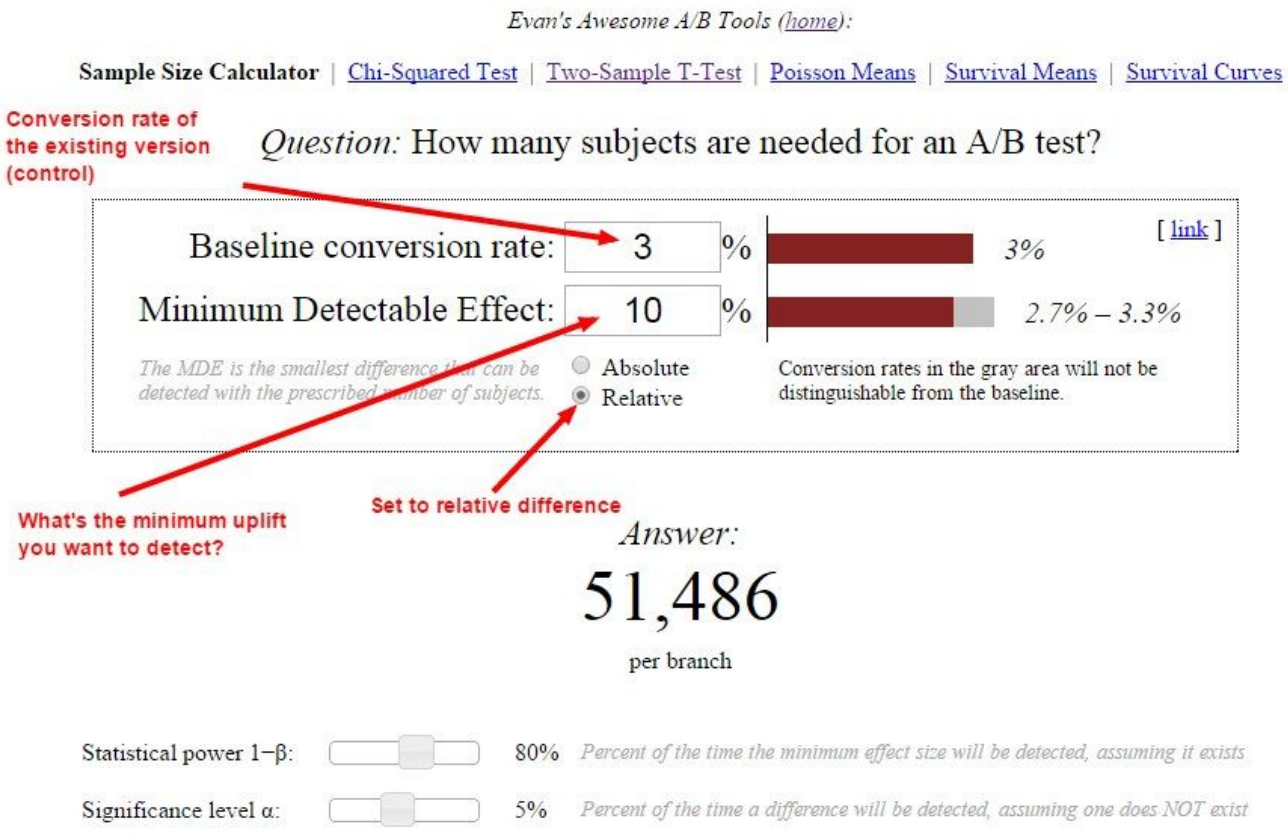
- ➢ ***This means if you've run 1.000 experiments and didn't control for repeat testing error in any way, a rate of successful positive experiments up to 25% might be explained by a false positive rate. But you'll see a temporary significant effect in around half of your experiments!***

So if you stop your test as soon as you see significance, there's a 50% chance it's a complete fluke. A coin toss. Totally kills the idea of testing in the first place.

Once he altered the experiment so that he would pre-determine the needed sample size in advance, only 51 experiments out of 1.000 were significant at 95%. So by checking the sample size we went from 531 winning tests to 51 winning tests.

How to pre-determine the needed sample size?

There are many great tools out there for that, like this one. Or here's how you would do it with Evan Miller's tool:



In this case we told the tool that we have a 3% conversion rate, and want to detect at least 10% uplift. The tool tells us that we need 51,486 visitors per variation before can look at the statistical significance levels and statistical power.

Magic numbers don't exist

What about the rules like X amount of conversions per variation?

Even though you might come across statements like "you need 100 conversions per variation to end the test" – there is no magical traffic or conversion number. It's slightly more complex than that.

It is never about how many conversions, it is about having enough data to validate based on representative samples and representative behavior.

And – if 100 conversions was the magic number, then big sites could end their tests just in minutes! That's silly. If you have a site that does 100,000 transactions per day, then 100 conversions can't possibly be a representative of overall traffic.

All significance reporting will be wrong if sample size is too small

I started a test for a client. 2 days in, these were the results:

Variations	Conversion Rate Range	Percentage Improvement	Chance to Beat Original	Conv/Visitors
Control	8.66% ±3%	-	-	11 / 127
Variation 1	0.91% ±1%	-89.50%	0%	1 / 110
Average Case	5.06% ±2%	-	-	12 / 237

The variation I built was losing bad – by more than 89%. Some tools would already call it and say statistical significance was 100%. The software I used (VWO) said Variation 1 has 0% chance to beat Control. My client was ready to call it quits.

However since the sample size here was too small (only a little over 100 visits per variation) I persisted and this is what it looked like 10 days later:

Variations	Conversion Rate Range	Percentage Improvement	Chance to Beat Original	Conv/Visitors	Action
Control	13.66% ±2%	-	-	87 / 637	
Variation 1	17.10% ±2%	+25.18%	95%	106 / 620	
Average Case	15.35% ±1%	-	-	193 / 1257	-

That's right – the variation that had 0% chance of beating control was now winning with 95% confidence.

Don't make conclusions based on a very small sample size.

Watch out for A/B testing tools "calling it early" – they always do. The reason is that statistical significance formulas make the critical assumption that the sample size was fixed in advance (and large enough). If instead of deciding ahead of time, "this experiment will collect exactly 10,000 observations," you say, "we'll run it until we see a significant difference," all the reported significance levels become meaningless.

Calculate the needed sample size in advance by using a sample size calculator such as this one to determine the needed sample size.

When looking at test results in your tools dashboard, always double check the total numbers. Always pay attention to the margin of error and sample size. Ignore results that have less than ~350 conversions per variation (unless B is *much* better than A).

Test results after the first day or two are most often gonna be completely different from the final results. “Peeking” at the data is OK as long as you can restrain yourself from stopping an experiment before it has run its course. That’s why it’s best not to give your boss or client access to real-time test results unless they’re made aware of this – or they might put pressure on you to stop the test.

Never let people stop tests early – it’s the #1 rookie mistake. Fight with everything you’ve got. If you can’t win, quit your job / fire the client. You’re just wasting your time.

Beware of case studies with low sample size

Most A/B testing case studies only publish relative increases. We got a 20% lift! 30% more signups! That’s very good, we want to know the relative difference. But can we trust these claims? Without knowing the absolute numbers, we can’t.

There are many reasons why someone doesn’t want to publish absolute numbers (fear of humiliation, fear of competition, overzealous legal department etc). I get it. There are a lot of case studies I’d like to publish, but my clients won’t allow it.

But the point remains – unless you can see test the duration, total sample size and conversion count per variation, you should remain skeptical. There’s a high chance they didn’t do it right, and the lift is imaginary.

How representative is the traffic in the test?

By running tests you include a sample of visitors in an experiment. You need to make sure that the sample is representative of your overall, regular traffic. So that the sample would behave just as your real buyers behave.

Some want to suddenly increase the sample size by sending a bunch of atypical traffic to the experiment. If your traffic is low, should you blast your email list, or temporarily buy traffic to get large enough sample size for the test?

- No.

In most cases you'd be falling victim to selection effect – you wrongly assume some portion of the traffic represents the totality of the traffic. You might increase conversion for that segment, but don't confuse that with an increase across segments.

Your test should run for 1 or better yet 2 business cycles, so it includes everything that going on:

- every day of the week (and tested one week at a time as your daily traffic can vary a lot),
- various different traffic sources (unless you want to personalize the experience for a dedicated source),
- your blog post and newsletter publishing schedule,
- people who visited your site, thought about it, and then came back 10 days later to buy it,
- any external event that might affect purchasing (e.g. pay day) and so on.

Regression to the mean

Some tests will win with 95% or higher confidence, but the uplift will vanish over time as the sample size gets larger. Experienced testers have seen it too many times. It makes you cringe.

The reason is that your "winner" was likely just a false positive. Of course the uplift doesn't stand up over time – that's because there was no uplift to begin with. This is a well-known phenomenon, called 'regression to the mean' by statisticians. This is common knowledge among statisticians but does not seem to be more widely known.

If you want to be sure of the result then always perform a second validation study to check if your results are valid.

Novelty effect

Some doubt whether this is real (and say it's actually regression to the mean), but "Novelty effect" is another explanation for why some tests seem to provider a winner that doesn't endure over time. The idea is that the uplift of the winner was not because of any actual improvement, but in response to increased interest in the new layout/offer/other test object.

Let's say you re-design the opt-in form on your website. Some returning visitors who've already opted in, might be drawn to it because its new – and opt-in again. That increases the conversion rate for the variation, but the performance is not sustained as the novelty wears off. So while you may get a winner shortly, the results fade as the sample size increases.

Have patience

Don't be discouraged by the sample sizes required – unless you have a very high traffic website, it's always going to take longer than you'd like. Rather be testing something slowly than to testing nothing at all. Every day without an active test is a day wasted.

Nine women can't give birth to a baby in 1 month. Your boss can say "test faster" all he wants, but it's not going to change the way math works. If you end tests before they're "cooked", you're screwing yourself, your company and wasting everyone's time. Don't do it.

Pay attention to the margin of error

Every test results will have a margin of error.

Example:

Here we have 3 variations.

- ➔ Original: 37.2% conversion rate
- ➔ V #1: 36.4% conversion rate
- ➔ V #2: 44% conversion rate

So if we didn't know to look at the absolute conversions, we could declare V #2 as the winner right?

Not so fast! If you look closer, every conversion rate is followed by a margin of error. In the case of V #2 its +/- 7.38%. So that means that it's not really 44%, but could be as low as 36.62% or as high as 51.38%! So now by seeing that the lower end for #2 is 36.62%, and the high for control is 37.2% + 6.79% = 43,99% – we can conclude that Control might very well be the actual winner. We don't know yet, we need more data! So the test has to keep going.

Conversion range can be described as the margin of error you're willing to accept. The smaller the conversion range – the more accurate your results will be. As a rule of thumb – if the 2 conversion ranges overlap (as they do in this example), you'll need to keep testing in order to get a valid result.

Craig Sullivan had this great write-up on error margins:

> The numbers are taking a guess at where the value might lie within and the testing software always shows a nice sweet point on a graph.
>
> However, it's not a precise point – it's a bell shaped curve probability region.
>
> When you see 3.5 +/- 0.1 it doesn't quite mean that it could be 3.6 or 3.4 – it means it's more likely to be in the middle (3.5), than out towards the edge but it's *possible*, just less so the further you get from the middle.
>
> I hate tests where this happens – because depending on how the results overlap, it means you're guessing about part or all of the result set. If there is a sizeable overlap, it means it's entirely possible that the results are winners for any of the overlapping ranges.

If just the edges of the intervals over lap (like 3.5 +/-0.4 and 2.7 +/- 0.4) then it's possible that the 2.7 one is the winner, just not as highly probable as the 3.5 one. With me so far?

OK – so if the ranges overlap lots (like 3.5 +/-0.2 and 3.4 +/- 0.2) then the bell shaped curve massively increases the likelihood that any 'winner' you draw from the experiment is not a result, just a guess and could turn out to be wrong.

You run the experiment and check the results every day and get a different 'winner'. If you forgot to check the data movements, you might drop by and sample the winner occasionally. It could be any one of them if they're still overlapping. Each time you look, it might be different orders showing.

This is why when one tests – you should always go for something that narrows in on the stuff that is shifting behaviour without overlaps. If I use a red button and that wins over a green button (and these two overlap heavily) – it just means there's not enough difference to be sure. It does not mean that red works better. This is the problem with any inference drawn from overlapping stuff. We get drawn to the contents we put in the test and look for confirmation from the data. Uh oh. Bad.

When you have a lot of overlaps, you really end up testing marginal things. You're doing what Google did with the shades of blue malarkey. If you keep the test running forever, every guess might turn out to be right along the way.

And the worst bit it is, if you pick an overlapping candidate as a winner, you may have picked the loser – next weeks data might have flipped things. If you pick an overlapping winner, it might go down in performance, leaving you to scramble to understand why or figure out what to do. I've done it for months on end and realised the futility of my stupidity in drawing useful stuff from overlapping results. You just end up chasing shadows, not being confident or really data driven.

Measure all the stuff that matters

What you measure is very important.

Measure final goals, not just page goals

You're running a test on an ecommerce product page. You're trying to get more people to add products to the cart with your test.

But if you're only measuring 'cart adds', you're doing it wrong. While this specific microconversion is an important metric to track, you need to track final purchases and revenue. Your treatment might get more people to add products to the cart, but it might also make less people complete the checkout.

Same thing with landing page optimization. Let's say you're working on a landing page that has an opt-in form, and is followed by a 4 step funnel. Even if you're trying to get more people to opt-in to a mailing list on the landing page, you need to be measuring to the end of the funnel. You want to find a variation that increases bottom of the funnel conversions.
So always measure until the final goal.

Measure to the money

Are you charging money on your website (as opposed to lead generation)? Make sure that you always track revenue. Number of transactions can go down while revenue goes up – so increase in absolute conversions are not the main goal. You're always after money!
So whenever possible, measure to the money.

Measure actions, not engagement

Engagement – the opposite of bounce – is an easy goal to track, but the least important one. Don't make any conclusions by just measuring the engagement rate. You need to measure the impact on business objectives.

Avoid overlapping tests when you can

You have lots of hypotheses, and you want to test them all. Can you save time by running multiple tests at the same time – one on the product page, one on the cart page, one on the home page?

Whenever possible, try to avoid it as it can skew the results. Technically there is a chance that it might not – but you can't know whether it does or not. So you can't be sure whether you can trust the data or not.

If you want to test a new version of several layouts at once—for instance product page, cart and checkout—you should use multi-page experiments designed for this purpose. That way people either see the new version for each page, or they see only the old ones.

Strategy One: Separate Tests

The most common, and easiest, st rategy is to just assume that the different tests don't really impact one another, and just run each test without bothering to account for the other test. In reality, this is often fine in many cases, especially if there is limited overlap.

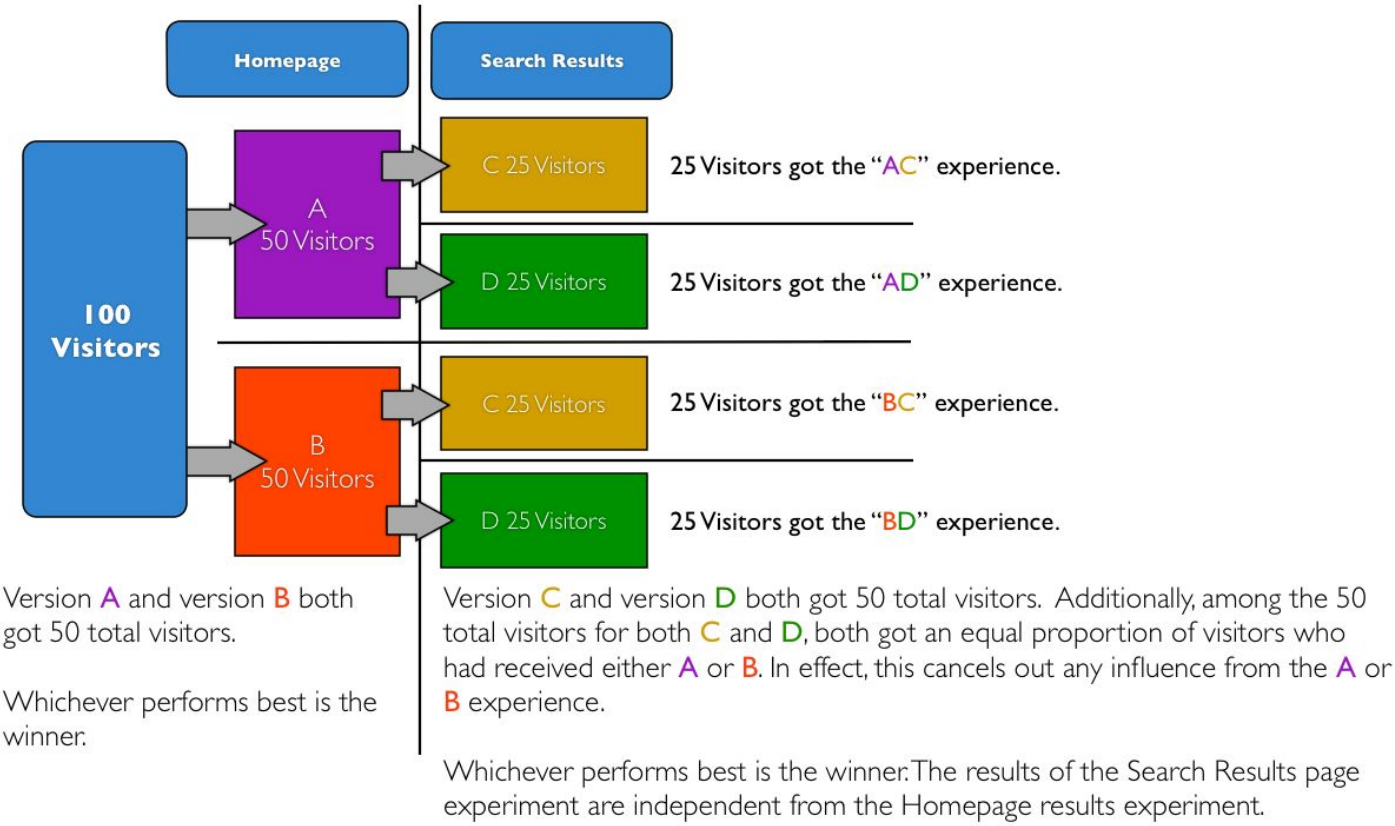
Condition: traffic needs to be distributed evenly.

Example:

You're running a test on your product page, you have product page A and product page B. Let's say that B is better and creates more motivation to buy the product. People who see variation B, and add product to cart, are taken onto a checkout page that you're running a test on. You have checkout page 1 and checkout 2. Now those people are distributed evenly: 50% go to checkout page 1 and 50% to checkout page 2. Since the traffic is split at random and evenly, both checkout pages get equal proportion of visitors who were positively impacted by product page B.

If traffic to checkout pages 1 and 2 were not distributed evenly, but say 25%/75% – then the results would get skewed. So when running tests with overlapping traffic, make sure the traffic is always split evenly.

Here's a graphy by Optimizely that explains the same:



Version A and version B both got 50 total visitors.

Whichever performs best is the winner.

Version C and version D both got 50 total visitors. Additionally, among the 50 total visitors for both C and D, both got an equal proportion of visitors who had received either A or B. In effect, this cancels out any influence from the A or B experience.

Whichever performs best is the winner. The results of the Search Results page experiment are independent from the Homepage results experiment.

While it may distort the absolute conversion rate, it doesn't matter since you care about relative conversion rate. You're trying to understand which variation performs best relative to other variations you're testing. As long as it's winning, keep going!

Strategy Two: The Multivariate Test.

We could combine both tests into one multivatiate test, with two decisions: Layout; Offer. This could work, but, once you start to think about it, maybe not the best way to go. For one, the test really only makes sense as a multivariate test if the user comes to the product section. Otherwise, they are never exposed to the product offer component of the test.

Also, it assumes that both the UX and Merchandising teams plan to running each test for the same amount of time. What to do if the UX team was only going to run the layout test for a week, but the merchandizing team was planning to run the Offer test for two weeks?

Strategy Three: Mutually Exclusive Tests

Rather then trying to force what are really two conceptually different tests into one multivariate test, we can instead run two mutually exclusive tests. This is to make sure users are assigned to just one of our tests.

Run tests for full weeks if your daily conversion rate varies

Let's say you have a high traffic site. You achieve 98% confidence and 250 conversions per variation in 3 days. Is the test done? Perhaps not!

We need to rule out seasonality, and to do that we need to test for full weeks. Did you start the test on Monday? Then you need to end it on a Monday as well. Why? Because your conversion rate can vary greatly depending on the day of the week.

So if you don't test a full week at a time, you're again skewing your results. Run a conversions per day of the week report on your site, see how much fluctuation there is. Here's an example:

Day of Week Name	Unique Visitors ↓	Ecommerce Conversion Rate	Transactions	Revenue
	536,048 % of Total: 100.00% (536,048)	3.41% Site Avg: 3.41% (0.00%)	26,362 % of Total: 100.00% (26,362)	$1,458,659.12 % of Total: 100.00% ($1,458,659.12)
1. Friday	126,658	3.37%	4,686	$240,615.79
2. Thursday	116,668	4.26%	5,507	$296,494.10
3. Tuesday	110,820	3.54%	4,318	$232,775.29
4. Saturday	105,714	2.43%	2,782	$151,139.37
5. Wednesday	96,170	3.55%	3,771	$213,462.54
6. Monday	86,675	3.46%	3,311	$204,184.84
7. Sunday	60,083	3.01%	1,987	$119,987.19

What do you see here? Thursdays make 2x more money than Saturdays and Sundays, and the conversion rate on Thursdays is almost 2x better than on a Saturday.

If we didn't test for full weeks, the results would be inaccurate. So this is what you must always do: run tests for 7 days at a time. If confidence is not achieved within the first 7 days, run it another 7 days. If it's not achieved with 14 days, run it another 7 days.

The only time when you can break this rule is when your historical data says with confidence that every single day the conversion rate is the same. But it's better to test 1 week at a time even then to rule out external factors and to increase the sample size.

Always pay attention to external factors

Is it Christmas? Your winning test during the holidays might not be a winner in January. If you have tests that win during shopping seasons like Christmas, you definitely want to run repeat tests on them once the shopping season is over. Are you doing a lot of TV advertising or running other massive campaigns? That may also skew your results. You need to be aware of what your company is doing.

External factors definitely affect your test results. When in doubt, run a follow-up test.

If you test more than 2 variations against control, and one of them is losing, don't just stop that variation

Let's say you set up a test where you have 4 variations + Control. After 5 days or so it's apparent that one or more of the variations are performing much worse than others. You might be tempted to stop sending traffic to those variations – but don't. You will alter the test composition if you do it.

What happens is that if you ramp up the traffic percentages in a test tool, or disable one or more variants as it's running, you're altering the proportions of the test.

There's a great explanation of it here in this paper – but best advice is when you're altering compositions of traffic in the test by removing variants or adjusting traffic percentages, beware of Simpsons Paradox.

Instead, weed out low performers in waves – stop the whole test, remove / disable lowperforming variations, and restart the test – leaving just the leading candidates + Control.

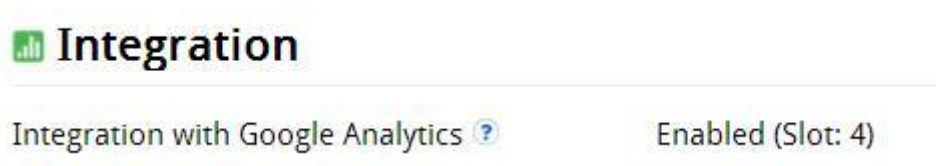
Integrate your testing tool with Google Analytics

Averages lie, always remember that. If A beats B by 10%, that's not the full picture. You need to segment the test data, that's where the insights lie.

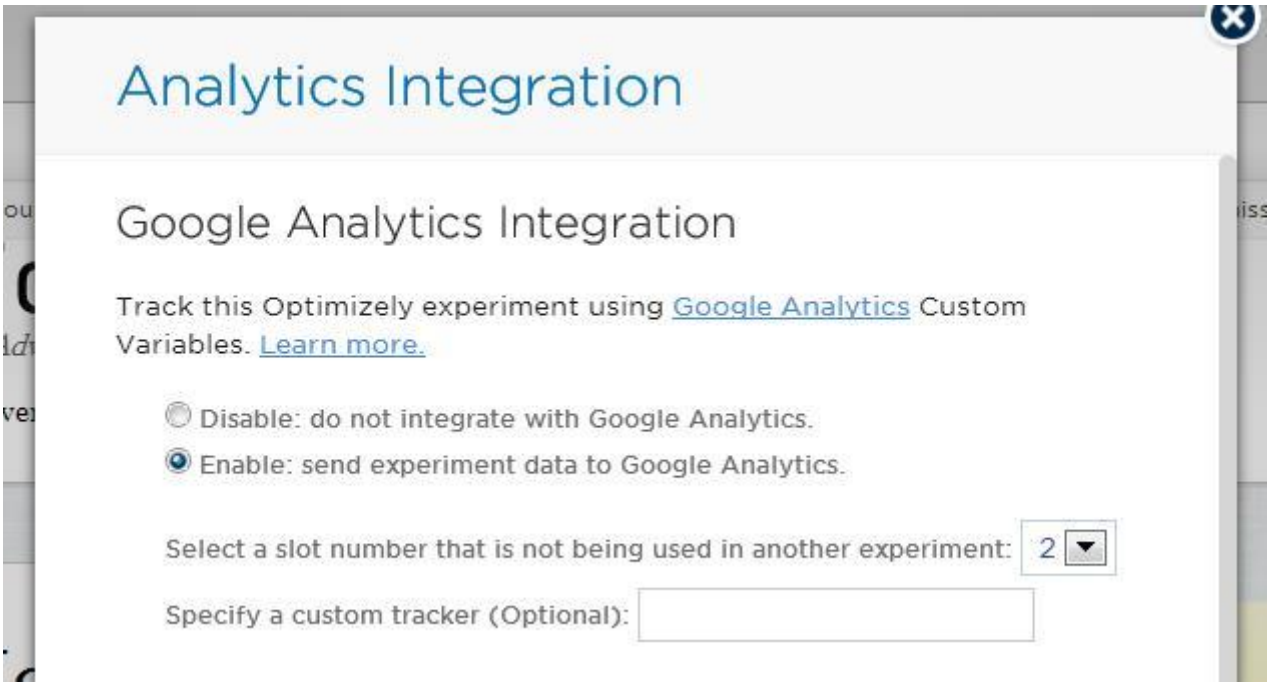
While Optimizely has some built-in segmentation of results, it's still no match to what you can do within Google Analytics. You need to send your test data to Google Analytics and segment it. Integration will also allow you to have two sources of performance data to triangulate or cross check with each other. If these don't line up proportionally or look biased to an analysts eye, this can pick up reporting issues before you've started your test.

If you use Visual Website Optimizer, they have a nice global setting for tests, so the integration is automatically turned on for each test you run.

Set it and forget it:



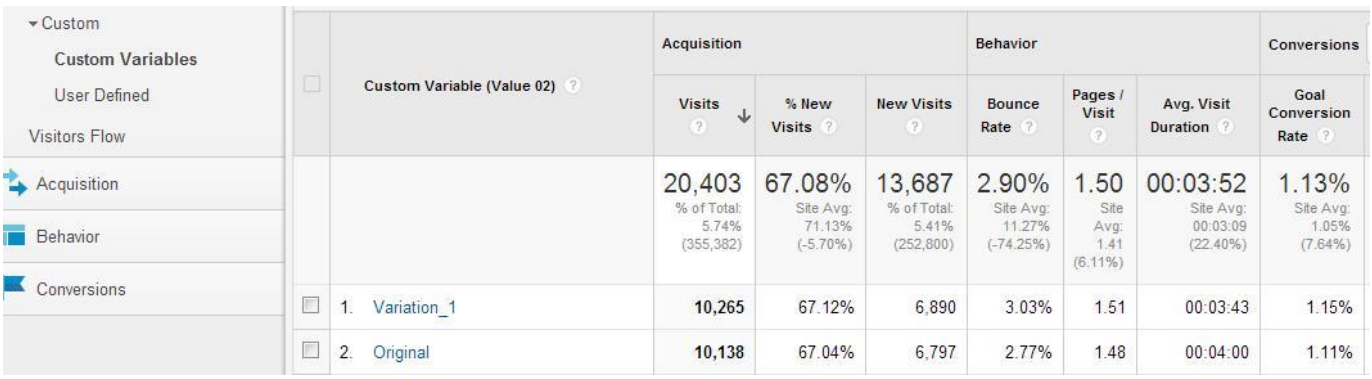
Optimizely makes you switch on the integration for each test separately.



Make sure all of your tests have a unique name, so they can tell them apart in Google Analytics:

Custom Variable (Key 1)	Acquisition		
	Visits ↓	% New Visits	New Visits
	32,246 % of Total: 1.05% (3,059,552)	24.42% Site Avg: 35.57% (-31.35%)	7,874 % of Total: 0.72% (1,088,298)
1. Optimizely_Checkout_Page_Red	31,943 (99.06%)	23.79%	7,599 (96.51%)
2. Optimizely_Improve_Account_C	295 (0.91%)	93.22%	275 (3.49%)

So what happens here is that they send the test info into Google Analytics as customvariables. You can run advanced segments and custom reports on it. It's super useful, and it's how you can actually learn from A/B tests (including losing and no-difference tests).



- Custom
  - Custom Variables
  - User Defined
  - Visitors Flow
- Acquisition
- Behavior
- Conversions

	Custom Variable (Value 02)	Acquisition			Behavior			Conversions
		Visits ↓	% New Visits	New Visits	Bounce Rate	Pages / Visit	Avg. Visit Duration	Goal Conversion Rate
		20,403 % of Total: 5.74% (355,382)	67.08% Site Avg: 71.13% (-5.70%)	13,687 % of Total: 5.41% (252,800)	2.90% Site Avg: 11.27% (-74.25%)	1.50 Site Avg: 1.41 (6.11%)	00:03:52 Site Avg: 00:03:09 (22.40%)	1.13% Site Avg: 1.05% (7.64%)
☐	1. Variation_1	**10,265**	67.12%	6,890	3.03%	1.51	00:03:43	1.15%
☐	2. Original	**10,138**	67.04%	6,797	2.77%	1.48	00:04:00	1.11%

But Monetate – which should be a class above the other two services, since it costs way more, is not even able to send custom reports. They can only send test data as events.

Top Events

- Event Category
- Event Action
- Event Label ▸

	Event Label	Total Events	% Total Events
1.	Control	76,151	50.15%
2.	Experiment	75,707	49.85%

view full report

So in order to get more useful data, create advanced segments for each variation and create a new segment based on the event label:

Top Events

- Event Category
- Event Action
- Event Label ▸

	Event Label	Total Events	% Total Events
1.	Control	76,151	50.15%
2.	Experiment	75,707	49.85%

view full report

And you can check whatever metrics in GA with a segment for each variation applied:

1.	New Visitor			
	All Visits	**546,738**	18,088	3.31%
	Monetate Control	**35,649**	2,720	7.63%
	Monetate Experiment	**34,580**	2,561	7.40%
2.	Returning Visitor			
	All Visits	**306,077**	69,818	9.18%
	Monetate Control	**27,955**	4,859	12.56%
	Monetate Experiment	**27,433**	4,728	12.36%

Bottom line: always send your test data to Google Analytics.