A/B TESTING: MISTAKES EVERYONE MAKES (AND HOW TO AVOID THEM)

KAREN HOPPER, SENIOR DATA STRATEGIST, M+R

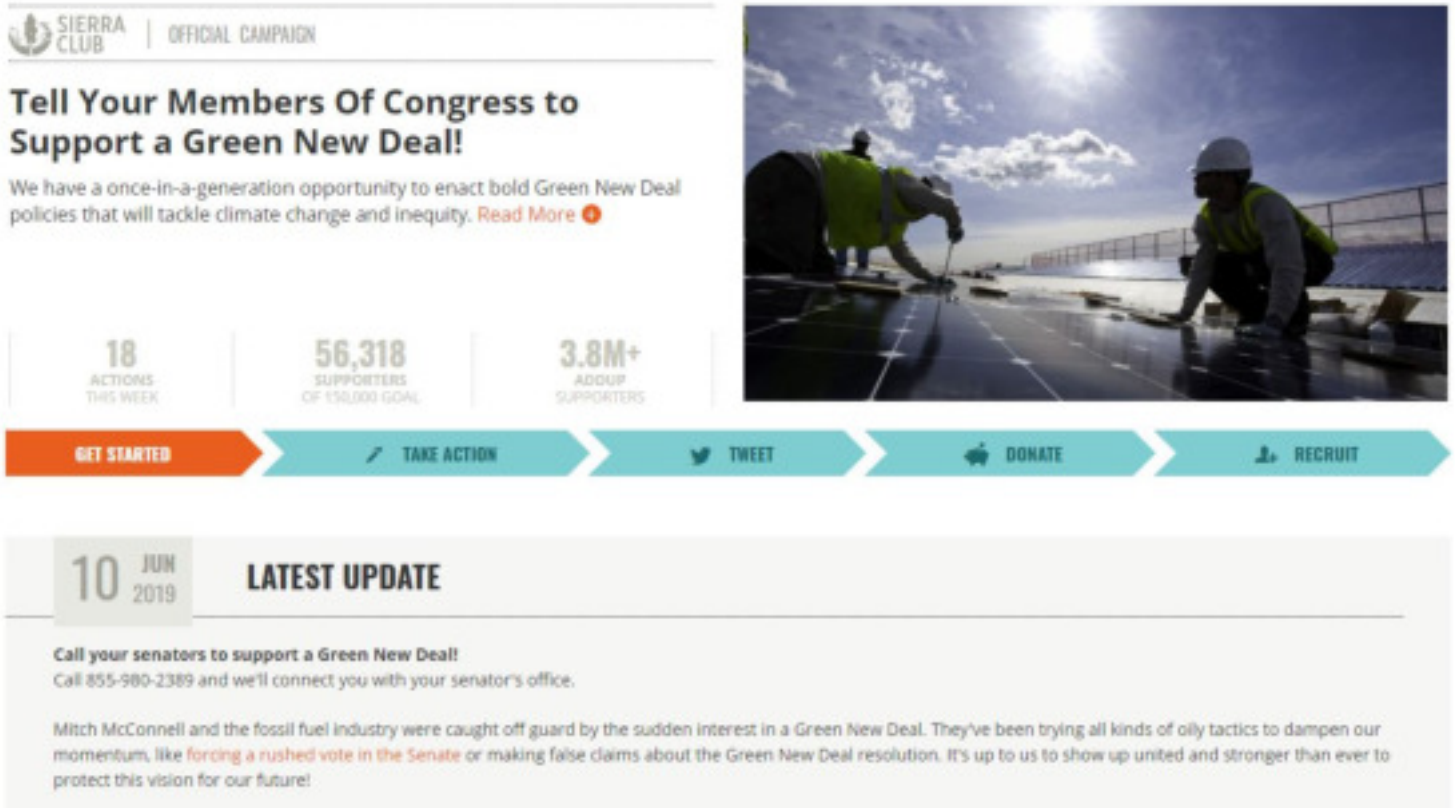
Hello! I'm Karen Hopper,
Senior Data Strategist,
M+R

We are communicators, marketers, fundraisers, and campaigners who unleash the power of people to do good.


Photo by Laura Ockel on Unsplash

Mistake 1: You’re Trying to Change Too Many Things at Once

The Kitchen Sink Problem

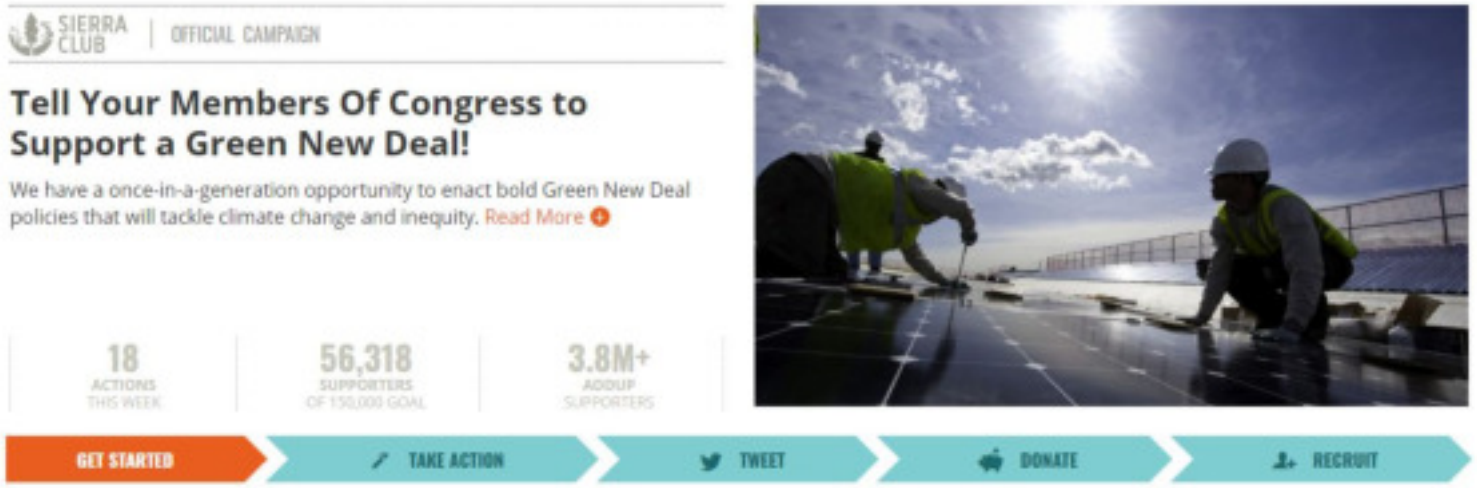


So which change had the impact?

Should we roll something out?

Mistake 2: You Don’t Have a Hypothesis For Your Test

Let’s come back to the action



The purpose of a hypothesis is to make sure we're thinking critically about why we want to test.

It also serves as a starting point for data collection, and to make sure our changes match up with our expectation of how user behavior might change.

Changing __(control)__ to __(variation)__ will improve __(test objective)__ because __(reason you think the change will help)__.

Test Description: Test an image including people on the action form against featuring animals only.

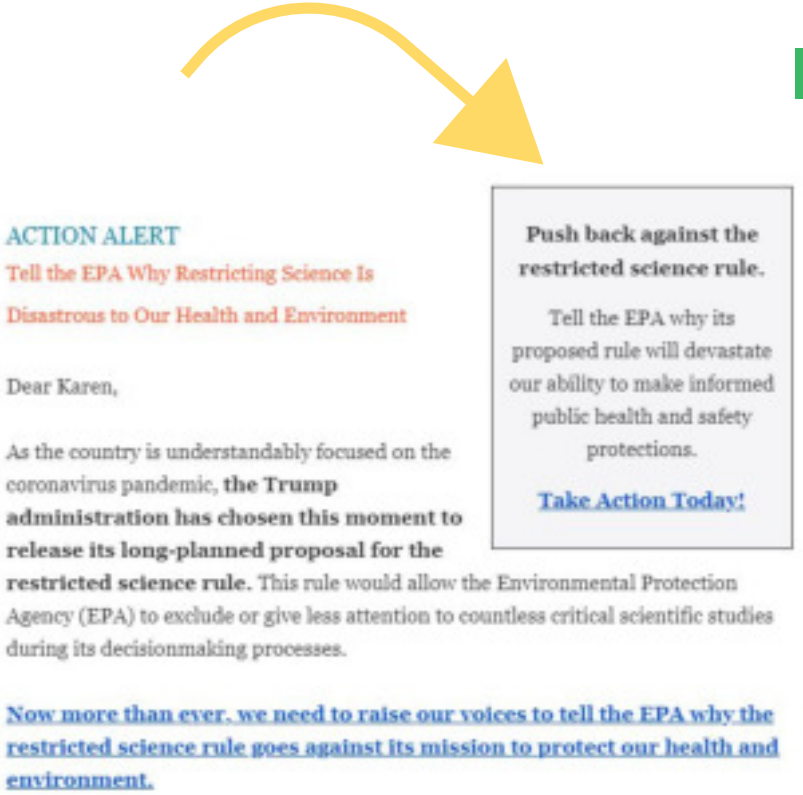
Hypothesis: Including people in the image instead of animals only will improve response rate because users will connect with the issue more naturally due to seeing people interact with the animals in the image.

Mistake 3: You Didn’t Randomize Your Audience

The tragic story of the call-out box format test.

How it went down:

- Unnamed international relief organization
- Sent 4 emails during a campaign testing the formatting of a call-out box to generate more click-throughs on their messages.
- Used ActionKit's built-in testing feature that allows you to split your audience for each send.
- Went to pull results to see which call-out box formatting won...



ACTION ALERT

Tell the EPA Why Restricting Science Is Disastrous to Our Health and Environment

Dear Karen,

As the country is understandably focused on the coronavirus pandemic, **the Trump administration has chosen this moment to release its long-planned proposal for the restricted science rule.** This rule would allow the Environmental Protection Agency (EPA) to exclude or give less attention to countless critical scientific studies during its decisionmaking processes.

Now more than ever, we need to raise our voices to tell the EPA why the restricted science rule goes against its mission to protect our health and environment.

Push back against the restricted science rule.

Tell the EPA why its proposed rule will devastate our ability to make informed public health and safety protections.

Take Action Today!

Womp womp

We couldn't use the results!

Because each message was re-randomized as it was sent, users saw both of the formats throughout the campaign and thus we couldn't combine results across appeals.

EOY 2016 Initial Results	...ients	Open Rate	CTR
EOY #8 (12/31)		**7.09%**	**0.17%**
Default		6.07%	0.20%
LYBNT		10.12%	0.80%
Propspect Control		7.21%	0.07%
Prospect Test	...,884	7.49%	0.08%
EOY #7 (12/31)	**466,524**	**7.28%**	**0.15%**
Default	169,667	7.07%	0.20%
LYBNT	34,066	10.06%	0.70%
Propspect Control	131,731	7.04%	0.05%
Prospect Test	...060	7.08%	0.05%
EOY #6 (12/31)		**8.31%**	**0.16%**
Default		8.33%	0.19%
LYBNT		12.30%	0.60%
Sustainer		15.17%	0.65%
Propspect Control	...219	7.31%	0.05%
Prospect Test	...31,551	7.44%	0.06%

How to Randomize

Read the fine print

Almost every CRM has a built-in testing function, but they're not all created equal.

Some re-randomize for every send or visit, which means that a user could be exposed to multiple test variations.

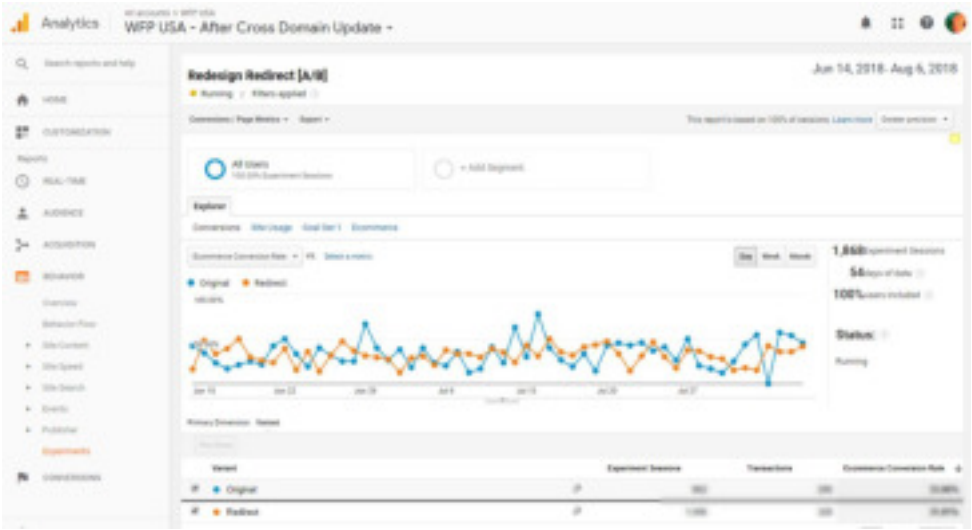
The one way to be certain is to manually split your file with excel. Another is to use a third party tool, such as Google Optimize.

- Export your list & open it in Excel
- Add a new column and paste **=rand()** in every cell in the column
- Sort by that column
- Split the list into two equal groups
- Import your two groups into your CRM

Mistake 4: You Check Your Test Too Many Times

It is the most tempting...

Every testing platform has a dashboard that offers a glance into how your test is doing *right now* - which makes it *so hard* to not snoop at the results and check it all the time.

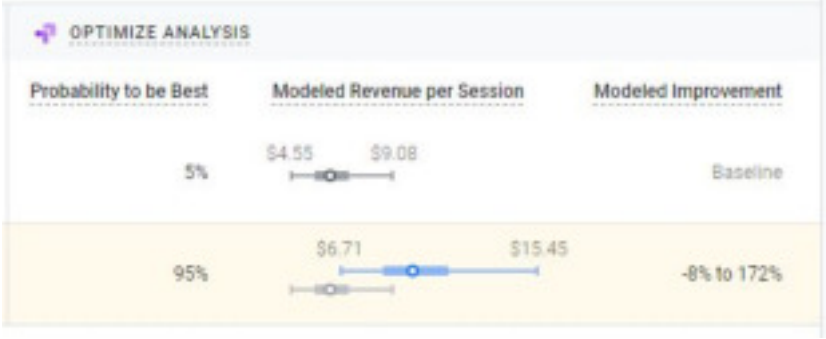


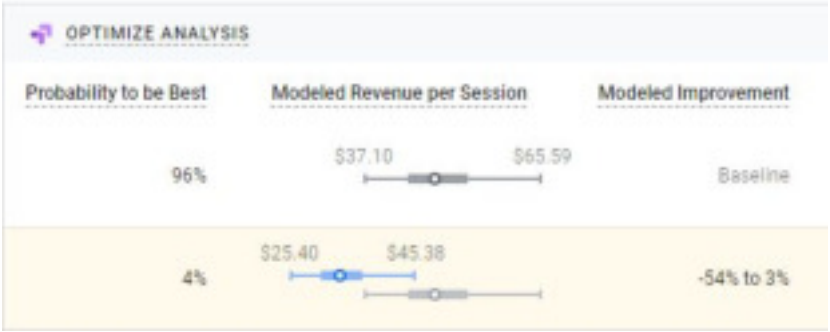
But every time you check it...

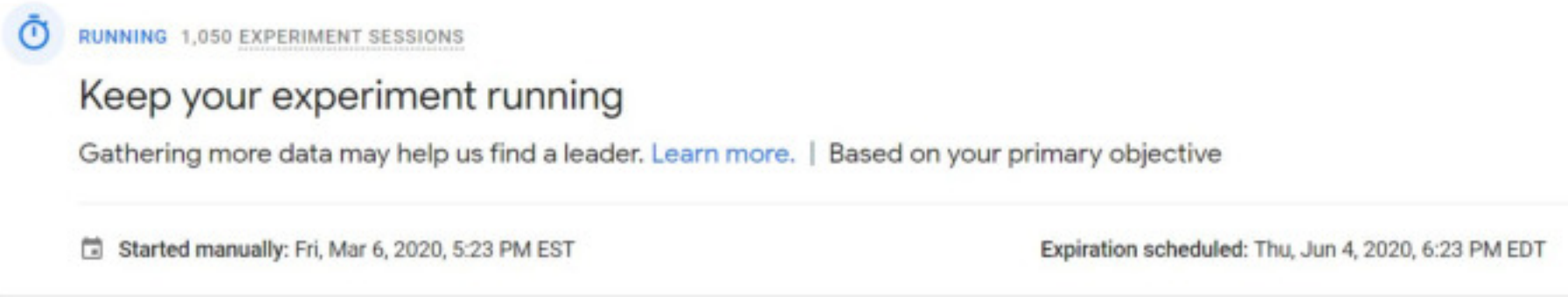
You're performing another statistical test on the results.

This is called "repeat significance testing" and can cause us to find the one moment in time that results *just happen to be* significant.

If we had stopped our experiment after seeing the "significant" results after a few days, we would have missed that the test variant was actually the *loser.*

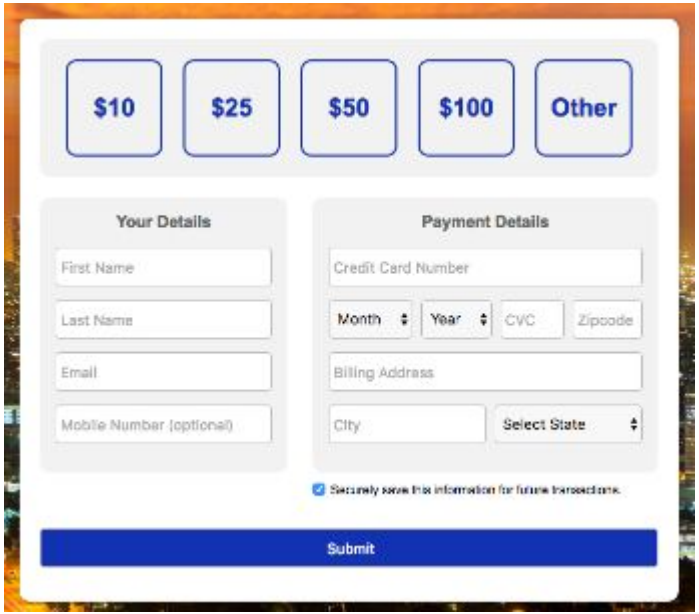





RUNNING 1,050 EXPERIMENT SESSIONS
Keep your experiment running
Gathering more data may help us find a leader. Learn more. | Based on your primary objective
Started manually: Fri, Mar 6, 2020, 5:23 PM EST
Expiration scheduled: Thu, Jun 4, 2020, 6:23 PM EDT

Mistake 5: Your Sample Size is Too Small

The Sad Donation Form Test



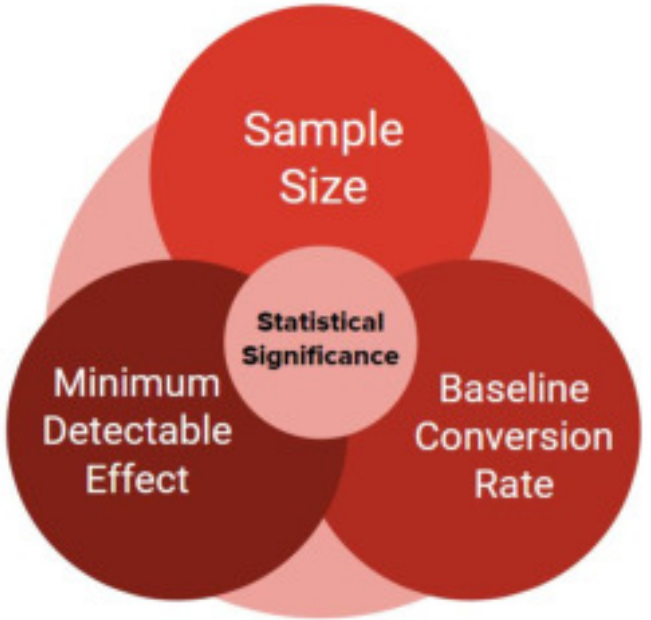
Only got 1,014 visitors in 12 weeks... 😱

Even a massive difference in performance doesn’t mean much if our audience isn’t big enough. Going from 3 conversions to 6 is a 100% increase, but it probably isn’t giving you a statistically significant result.

Feel the POWER

Power is a measure of how well you can distinguish the difference you're seeing in your experiment vs random chance.

Without the enough power - based on how much change you expect to see, how many people you're testing on, and what the initial rate you're starting from - it's unlikely that you'll see statistical significance.



Needed Sample Size (per variant):

		Improvement		
Baseline Conversion Rate		5%	10%	25%
	1%	458,900	101,600	13,000
	5%	69,500	15,000	1,800
	10%	29,200	6,200	700
	25%	8,100	1,700	200

https://www.optimizely.com/sample-size-calculator/

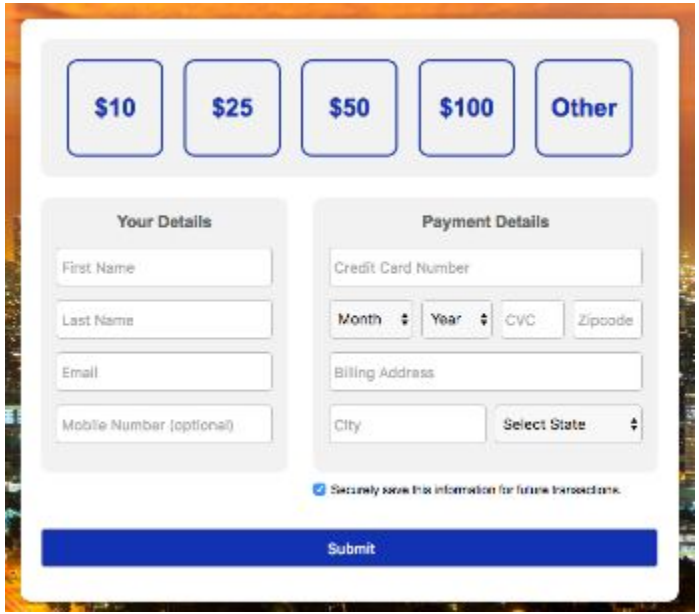
What if I need a bigger sample size?

Can you...

- Run the test over the course of an entire campaign instead of in just one email?
- Add additional segments or channels to your experiment?
- Remove one or more variants?
- Or.... Can you test something where you think you'll see a bigger impact?

Mistake 6: The Thing You’re Changing Is Too Small

Remember the Sad Donation Form Test?



I see this a lot!

Testing is risky so...

- Let's just test the color on the links in this email.
- What happens if we remove one word from our CTA?
- Which photo is better, this one or that one (of the same animal)?
- Let's just test subject lines in this email campaign
- What if we changed our donate button color?

Needed Sample Size (per variant):

		Improvement		
Baseline Conversion Rate		5%	10%	25%
	1%	458,900	101,600	13,000
	5%	69,500	15,000	1,800
	10%	29,200	6,200	700
	25%	8,100	1,700	200

https://www.optimizely.com/sample-size-calculator/

Go big(ish) or go home.

Mistake 7: You Aren’t Relying on Statistical Tests to Determine a Winner

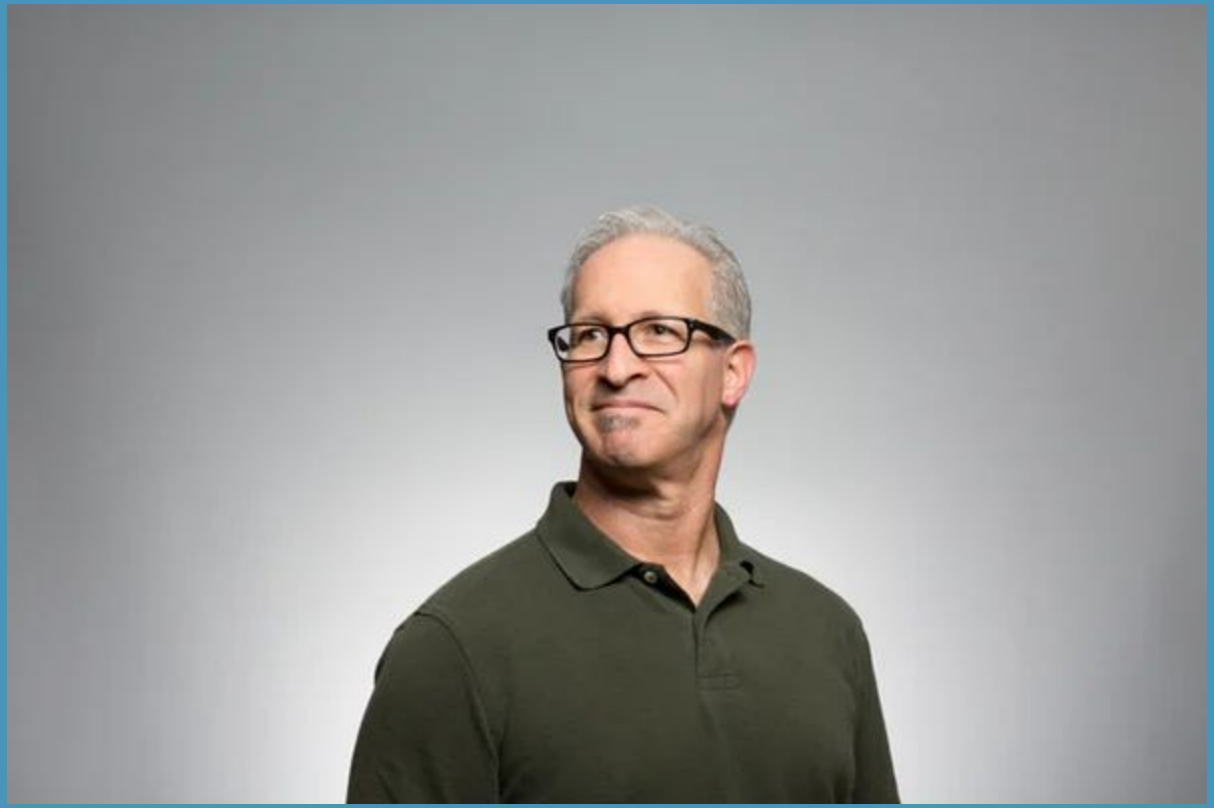


Photo by Foto Sushi on Unsplash


NEVER SAY A VARIANT WON IF THERE
WAS NO STATISTICAL SIGNIFICANCE.

Statistical Significance

95% Confidence

This means that if we repeat the experiment over and over again to the same group of people under the same conditions, 19 out of 20 times, we can expect the same outcome will occur.

There is some risk that random chance will favor one version over another, but by using the 95% confidence interval, we are reducing our risk that we are picking the wrong condition.

But what if you need to choose one and there's no "winner"?

Decide before you test!

Usually, we sit down with our team before the test runs and have a discussion about what to do if the control wins, the test wins, or if it's a draw. Sometimes it's good that there's no difference (like during a rebrand)! Other times, it just means we can pick the one we like.

Just make sure you're being up front about your decision making - and that you aren't inauthentic about declaring a winner.


Photo by Laura Ockel on Unsplash

A big fat caveat:

These learnings only apply if you're setting out to do this kind of controlled testing. Some things, like Facebook Ad Optimization, Unbounce Smart Traffic, and other tools that are "testing" rather than TESTING will give you winners based on more than just a statistical test.

QUESTIONS?

Additional Resources!

For testing fun

How long should your test run?

Run your numbers through the Sample Size calculator first.

Common mistakes (and how to avoid them) if you want a written recap of this webinar.

Need a tool for measuring results? M+R has that too!

Every test should have a plan - here's a template from Optimizely that I love to use.


M+R