How to fail: Optimize learnings from every test that you run

Ronnie Cheung
Senior Strategy Consultant
Optimizely

Laura Shelton
Optimisation Manager
Channel 4

Marcello Pasqualucci
Head of Experimentation
SKY

Charlotte Golding
Head of Optimisation
Virgin Media

Let’s talk about failures in testing

Can you guess what is the average win rate across all experiments ran by Optimizely? 10%

Common “Failures”

Statistically Significant, *but Negative*

Inconclusive

Flawed Test Design

Why does it suck so much?

The feeling of going back to square one

Wasted time and resources

You weren’t able to prove someone wrong

Type of “Failure” #1

Statistically Significant, *but Negative*

How people read statistically significant negative test results

How people SHOULD read statistically significant negative test results

Every negative result is still a win

Validated a feature / an experience
Instead of going straight build, you were able to identify that this variation will not generate an positive improvement

Learned about your customers / users
An underperforming variation tells you something about your users' preferences, which you didn't know about

Concluded results with confidence
Reaching statistical significance is typically a sign of a well designed, sufficiently powered test.

We should celebrate 🎉

What were the interesting things you learned about your users through experiments with stat. negative results?

SUCCESSFUL

Register

On your marks, get set... bake!

Opt in to hear about:

- The latest from inside the Bake Off tent and a showstopping recipe of the week in your inbox
- Returning shows you love
- Exclusive rewards and competitions
- And other delicious All 4 Exclusives

Opt in and register

Register

By continuing, you confirm your date of birth is correct, you confirm you have read our Privacy Policy, and you accept the Terms and Conditions and Channel 4's use of cookies as set out in our Cookies Policy.

SUCCESSFUL

Register

Don't miss new comedy...

Opt in to hear about:

- Our latest UK comedy
- Hit US sitcoms
- Brand-new and exclusive shows
- Competitions and rewards

Opt in and register

Register

By continuing, you confirm your date of birth is correct, you confirm you have read our Privacy Policy, and you accept the Terms and Conditions and Channel 4's use of cookies as set out in our Cookies Policy.

UNSUCCESSFUL

Register

Don't miss new entertainment...

Opt in to hear about:

- Our top entertainment
- From MAFS and MiC to Gogglebox and Bake Off
- Brand-new and exclusive shows
- Competitions and rewards

Opt in and register

Register

By continuing, you confirm your date of birth is correct, you confirm you have read our Privacy Policy, and you accept the Terms and Conditions and Channel 4's use of cookies as set out in our Cookies Policy.

UNSUCCESSFUL

Register

Don't miss new entertainment...

Opt in to hear about:

- Our highest-rating entertainment
- Reality hits from the UK, US and Australia
- Brand-new and exclusive shows
- Competitions and rewards

Opt in and register

Register

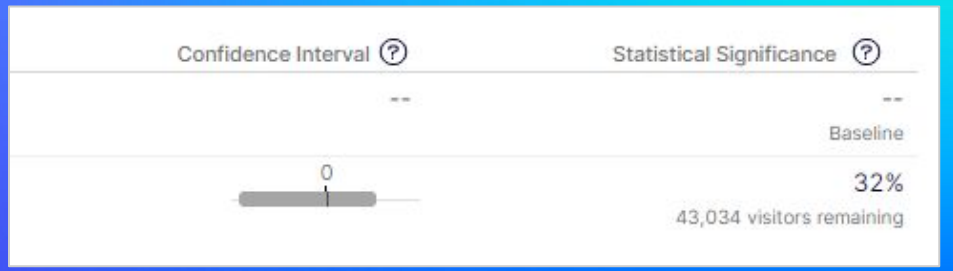
By continuing, you confirm your date of birth is correct, you confirm you have read our Privacy Policy, and you accept the Terms and Conditions and Channel 4's use of cookies as set out in our Cookies Policy.

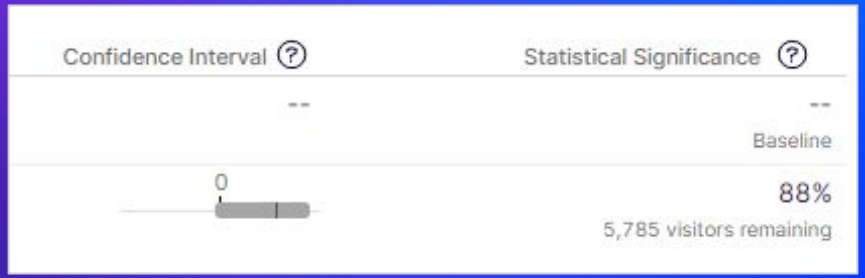
What advice would you give if someone receives a statistically negative results, in terms of the actions to take?

Type of “Failure” #2

Inconclusive

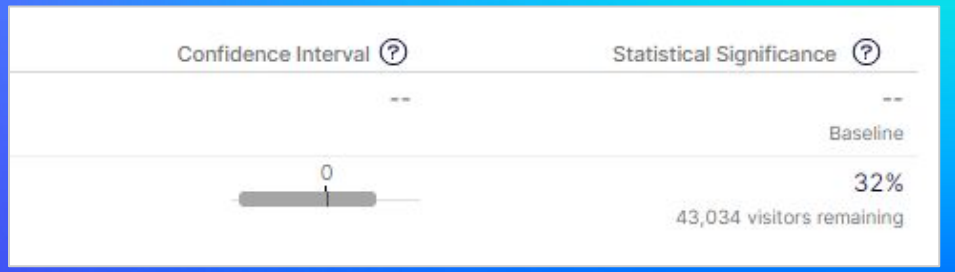
Not all inconclusive are the same



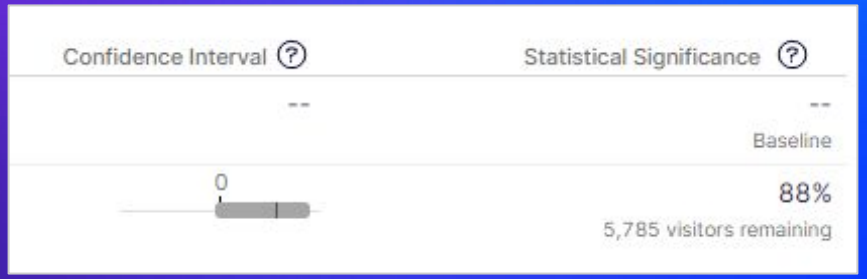


What does it mean?

No impact being seen in treatment



Some impact being seen in treatment, low confidence



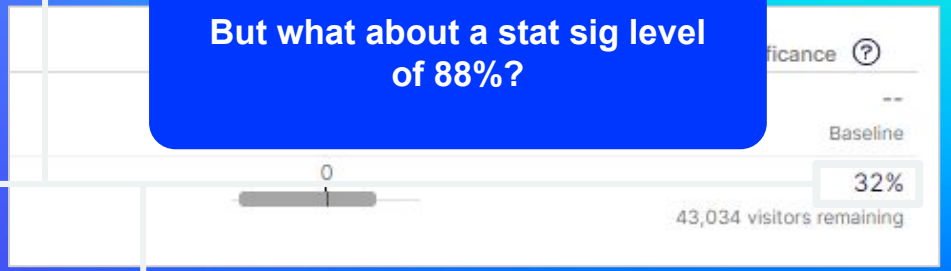
Impact being seen in treatment, but not statistically significant

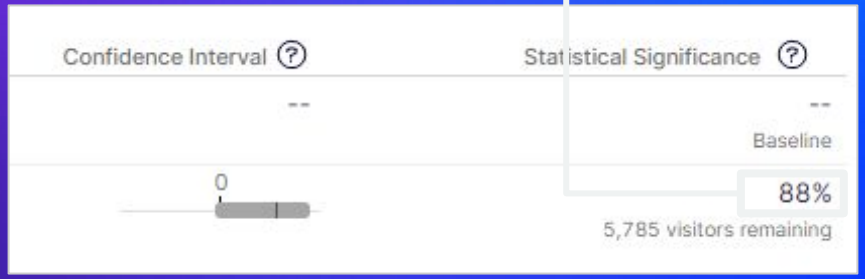
What should we do?

Let's say we accept stat sig of 90%.

A stat sig level of <1% or 32% are **weak evidence** of a real effect from the treatment.

But what about a stat sig level of 88%?





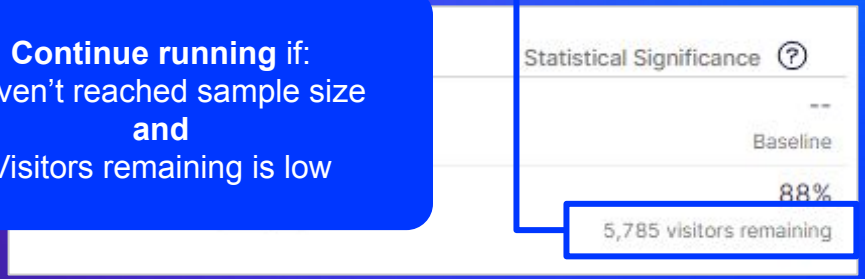
To run or to pause?

How many visitors remaining?

Compare visitors remaining to the estimated sample size in your pre-test calculations.

Declare inconclusive if:
Reached estimated sample size
and
Visitors remaining is high

Continue running if:
Haven't reached sample size
and
Visitors remaining is low



Day 9

CX-2 Click PRIMARY METRIC

Edit

Unique conversions per visitor for **CX-2 Click** event

	Unique Conversions Visitors	Conversion Rate	Improvement	Confidence Interval	Statistical Significance
Original	34 9,368	0.36%	--	--	-- Baseline
Variation #1	297 9,444	3.14%	+766.5%	0	29% 328 visitors remaining

Day 20

CX-2 Click PRIMARY METRIC

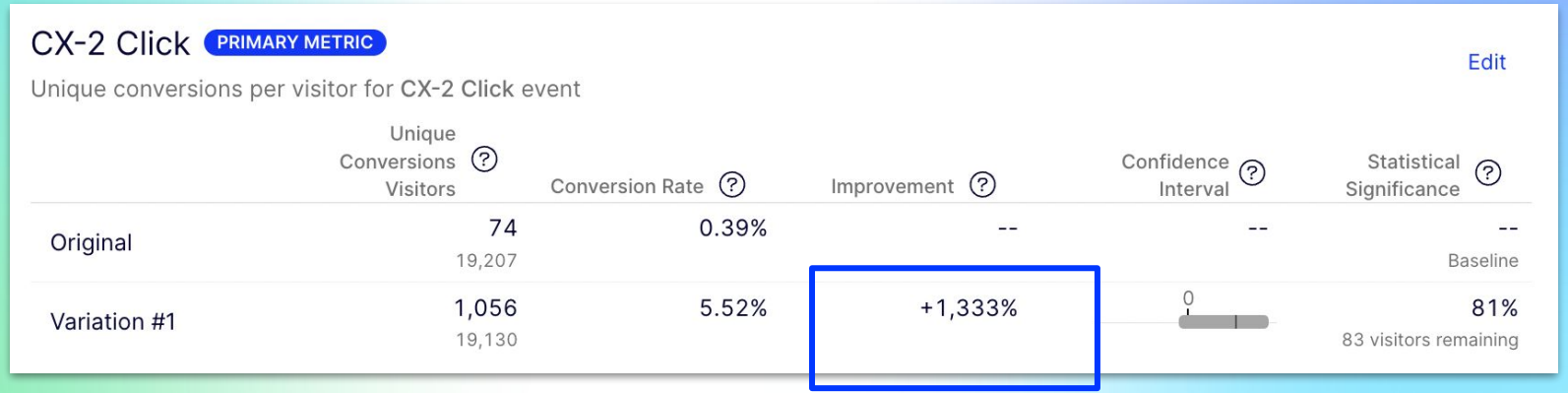
Edit

Unique conversions per visitor for **CX-2 Click** event

	Unique Conversions Visitors	Conversion Rate	Improvement	Confidence Interval	Statistical Significance
Original	74 19,207	0.39%	--	--	-- Baseline
Variation #1	1,056 19,130	5.52%	+1,333%	0	81% 83 visitors remaining

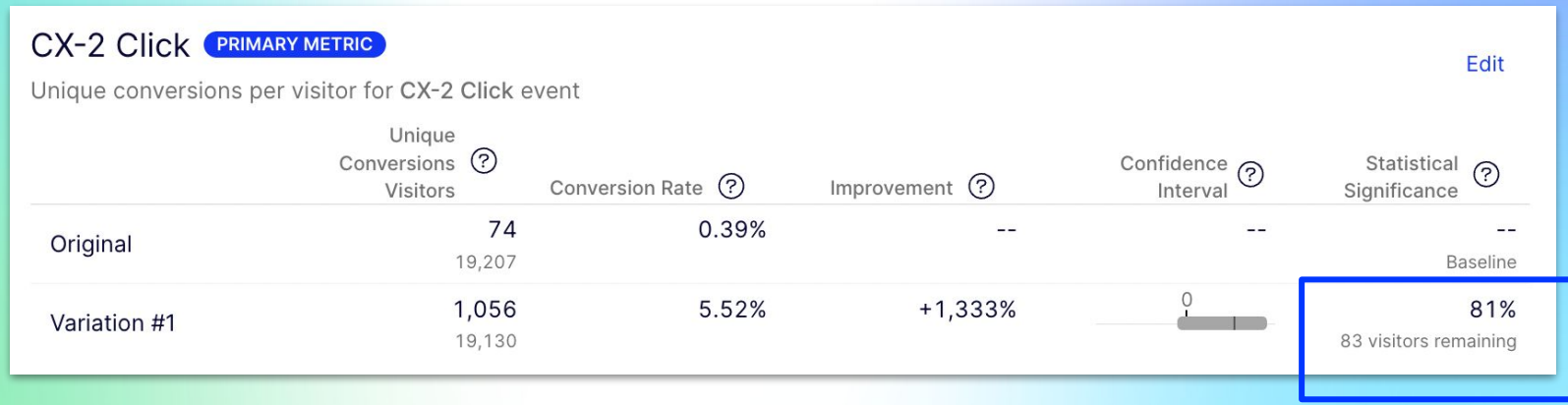
“Variation is showing +1,333%, why is it still not reaching stat sig?”

Your baseline conversion matters. 74 conversions mean there is not enough sample data from the control to validate your results.



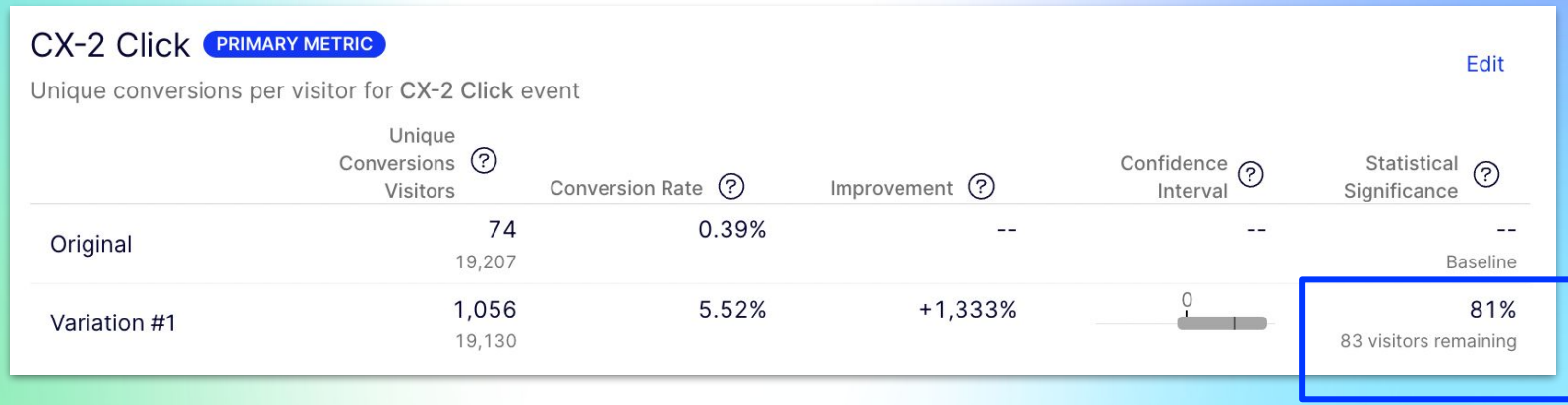
That being said, statistical significance is about risk tolerance.

If you are confident that the variation performs better, then there's no reason why you shouldn't implement this.



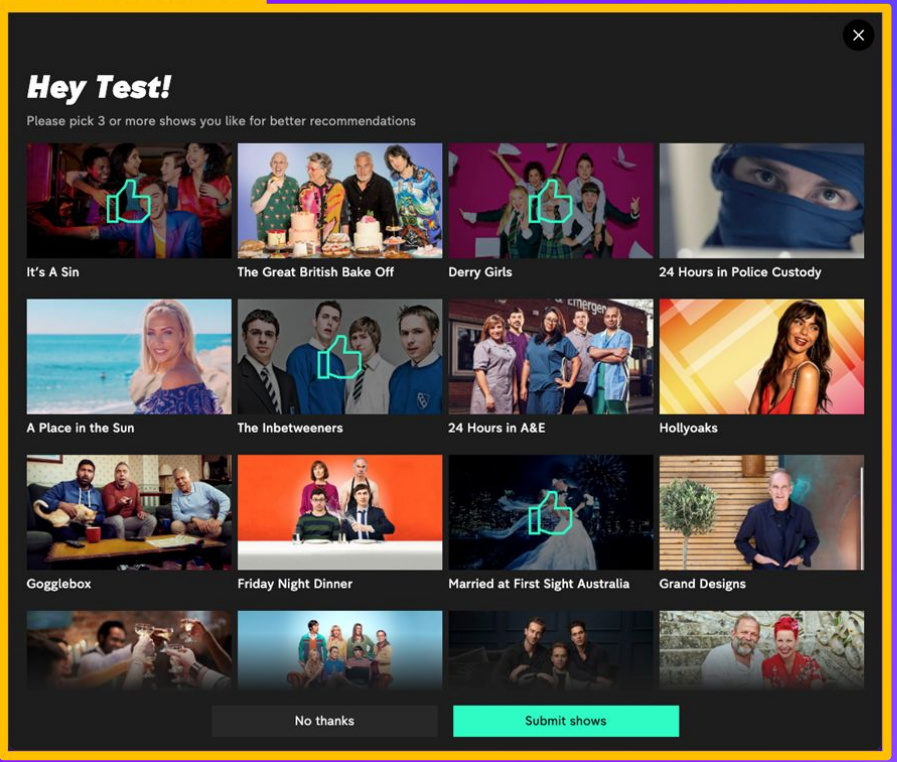
Think contextually about your user journey.

You may have a high **risk tolerance** towards a test that's running on the final checkout page, versus a test that's running on a confirmation page.

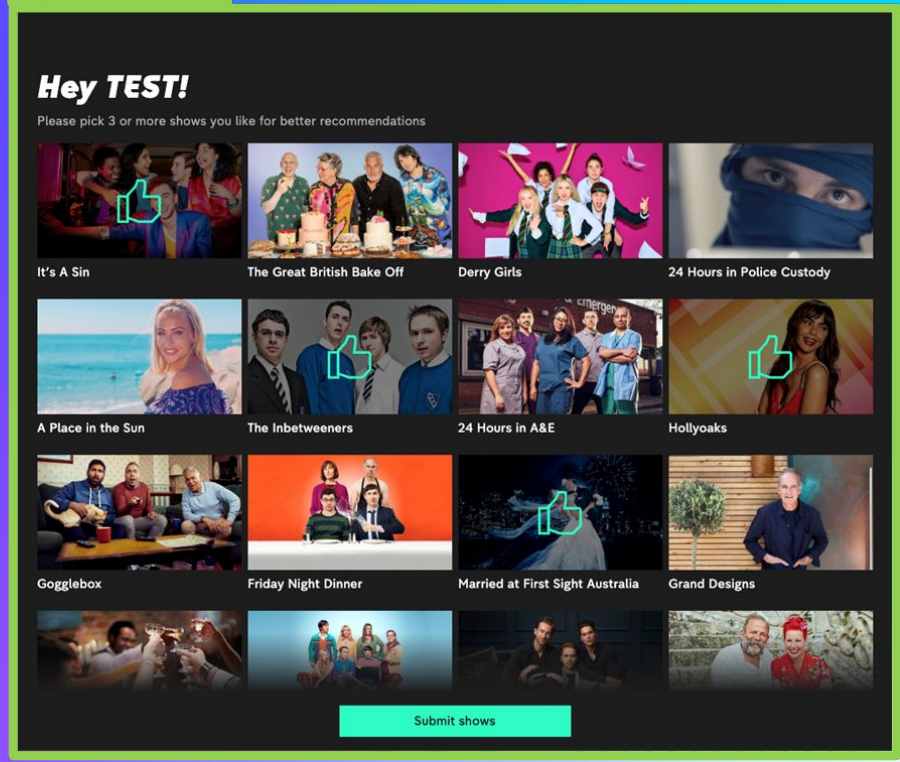


Do you have any examples where you saw the opportunities from an inconclusive experiment?

INCONCLUSIVE

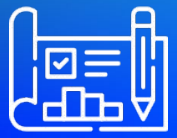


SUCCESSFUL



What advice would you give if a team keeps getting inconclusive results?

Type of “Failure” #3



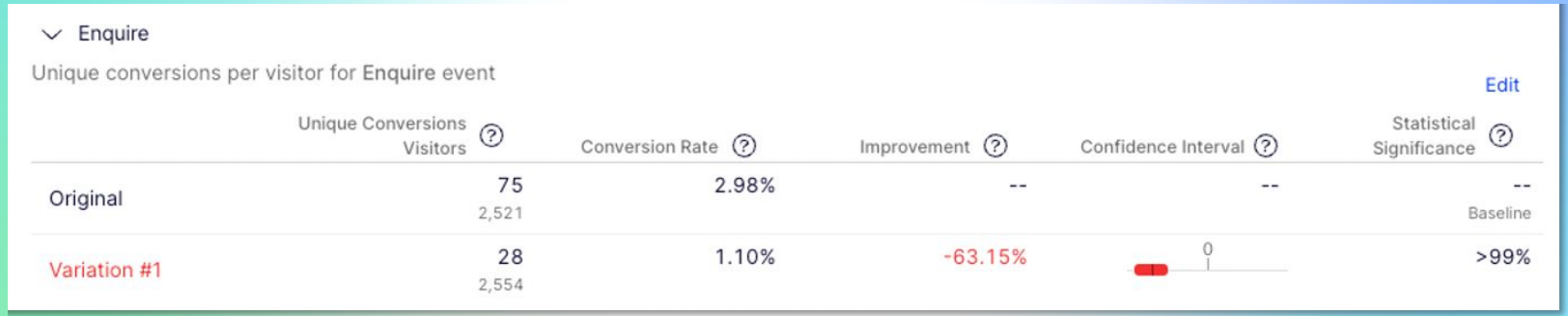
Flawed Test Design

Mistakes we see again and again…

No qual/quant. backing

This prevents you from being able to identify and address your user problems

Testing that home buyers in consideration phase prefer “softer copy” and removed the home loan interest rate.



Enquire

Unique conversions per visitor for Enquire event

Edit

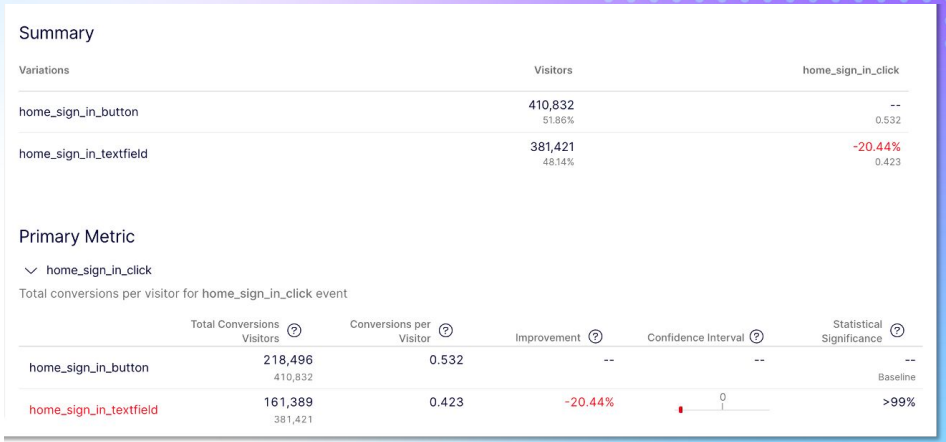
	Unique Conversions Visitors	Conversion Rate	Improvement	Confidence Interval	Statistical Significance
Original	75 2,521	2.98%	--	--	-- Baseline
Variation #1	28 2,554	1.10%	-63.15%	0	>99%

Not assigning the right metrics

Your primary metric should also be the direct action that a user can take.

This experiment only included one metric in the set up.

The variation was unsuccessful, but we couldn't understand why without the ability to evaluate other behavioural metrics.



Summary

Variations	Visitors	home_sign_in_click
home_sign_in_button	410,832 51.86%	-- 0.532
home_sign_in_textfield	381,421 48.14%	-20.44% 0.423

Primary Metric

home_sign_in_click

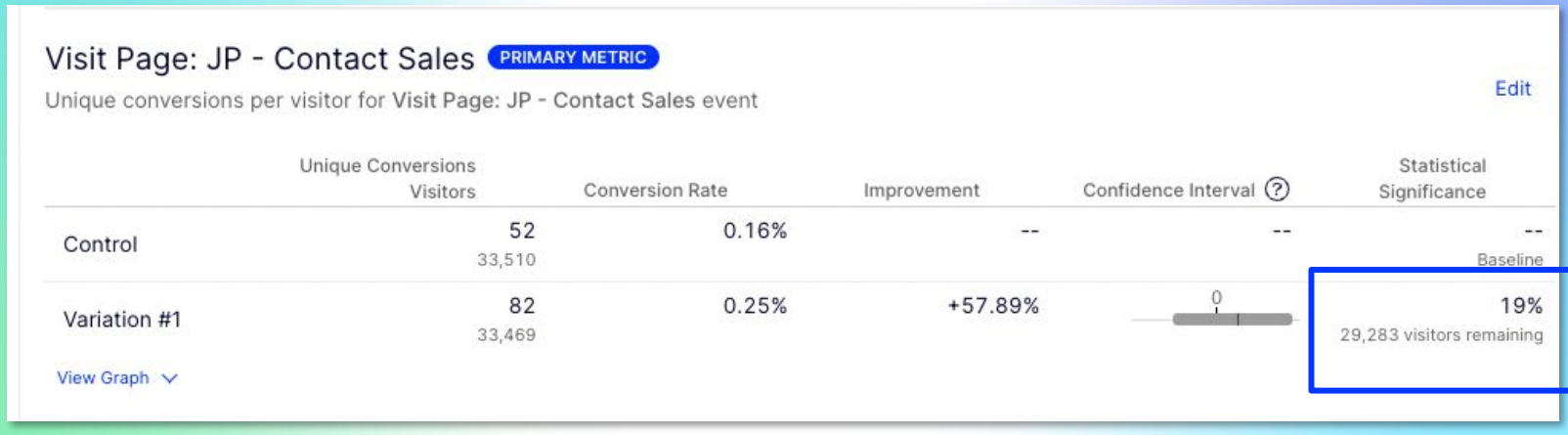
Total conversions per visitor for home_sign_in_click event

	Total Conversions Visitors	Conversions per Visitor	Improvement	Confidence Interval	Statistical Significance
home_sign_in_button	218,496 410,832	0.532	--	--	-- Baseline
home_sign_in_textfield	161,389 381,421	0.423	-20.44%	0	>99%

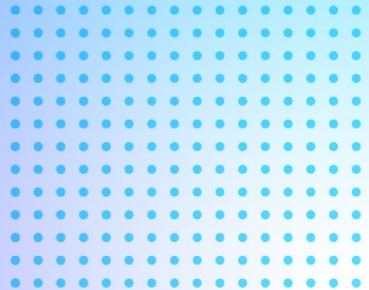
Not doing the pre-test calculations

You start launching experiment(s), only to realise none of them are going to reach stat sig.

This CTA colour change test has been running for almost 10 weeks.



What were some of the painful lessons you learned in the past about test designs?




Cohort of users
BAU users to promote show 1
BAU users to promote show 2
BAU users to promote show 3
Users who don't fit any show
Control group
Control group
Control group
Show 1
Recommender chosen
Show 2
Recommender chosen
Show 3
Recommender chosen
BAU users who qualify for show 1, 2, 3, 4 or 5
Control % TBC
BAU group
Recommender group
BAU users to promote show 1
BAU users to promote show 2
BAU users to promote show 3
BAU users to promote show 4
Recommender show 1
Recommender show 2
Recommender show 3
Recommender show 4
Cohort of users
Control group
BAU targeted group
Recommender targeted group
Users who don't fit any campaign
Campaign 1
Campaign 2
Campaign 3
BAU users identified for Campaign 1
BAU users identified for Campaign 2
BAU users identified for Campaign 3

How do you make sure all teams follow best practices when setting up experiments?

What have we learned?

Every (stat. sig) negative result is still a win.

Not all inconclusive are the same.

A proper test design can improve your win rate and conclusive rate.

Thank you

Laura Shelton
Optimisation Manager
Channel 4

Marcello Pasqualucci
Head of Experimentation
SKY

Charlotte Golding
Head of Optimisation
Virgin Media