GrowthBook

Trustworthy A/B Testing: Red Flags Before You Ship

Ronny Kohavi
Ex-VP & Technical Fellow Airbnb, Microsoft, & Maven Instructor

Luke Sonnet
Head of Experimentation GrowthBook

Housekeeping

Session Recording

This session is being recorded, we'll send it out via email

Zoom Chat

Use the Zoom chat for quick comments or clarifying questions

Questions?

Ask questions and vote in the Q&A panel anytime; we'll answer some of the highly-voted questions at the end

Help Us Improve

A quick survey will pop up at the end

Agenda

1. **Anatomy of an Untrustworthy A/B Test: The Four Red Flags**

   Dissecting a published 44% lift
   Power calculations, p-values, SRM, Twyman's Law

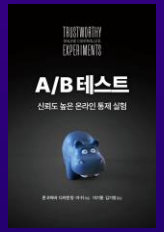
2. **Silent Killers: Peeking & Multiple Comparisons**

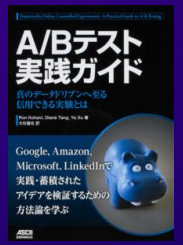
   How they compound your error rate and what to do about it

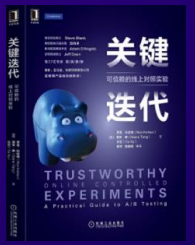
3. **Raffle and Q&A**

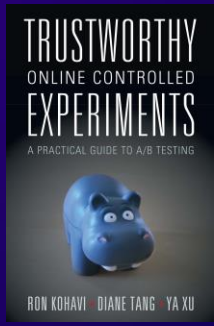
Who is Ronny Kohavi (and my biases)

- PhD in Machine Learning from Stanford
- Director of Data Mining and Personalization, Amazon
  - One of my teams was Weblab, Amazon's experimentation system
- Technical Fellow and Corporate VP, Microsoft (14+ years)
  - Led the Analysis & Experimentation team
  - Large-scale: 100+ new treatments started every workday
- Vice President and Technical Fellow, Airbnb
  - Relevance, personalization, and experimentation
  - Co-authored the HiPPO book (https://experimentguide.com)
    Trustworthy Online Controlled Experiments: A Guide to A/B Testing
    - Available in English, Chinese, Japanese, Korean, Russian, Polish
  - Now consulting and teaching quarterly





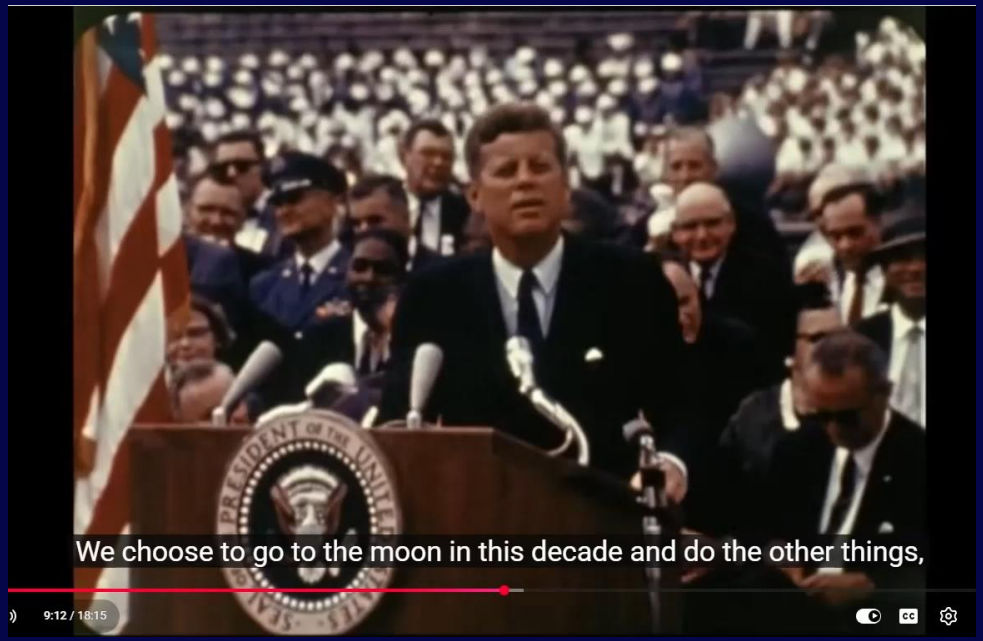




Trustworthy Results are Harder Than You Think

- President John F. Kennedy said the following on Sept 12, 1962

  We choose to go to the moon in this decade and do the other things,
  not because they are easy, but because they are hard…



Trustworthy Results are Harder Than You Think

- President John F. Kennedy said the following on Sept 12, 1962
  We choose to go to the moon in this decade and do the other things,
  not because they are easy, but because they are hard…

- I see many organizations thinking about the following variant
  We do this not because it is easy
  but because…
  we thought it would be easy ☺
- A/B tests are the gold standard in science, but trustworthy A/B tests require effort. Back in 2010, we wrote:
  Getting numbers is easy;
  getting numbers you can trust is hard

Motivating Example

Article published an A/B test claiming

- Treatment had 44.8% conversion lift (from 55.7% to 80.6%)
  The details of the treatment are not important for this analysis,
  but it was buzzword compliant using AI
- 99% confidence, which the author wrote “means we can be 99% certain the result reflects a genuine difference between the two pages, not a statistical accident.”
- Visitors split evenly across both variants with ~54,000 people per variant
- Conclusion: “AI smashed it right out of the park!”

We will now show why the result is not trustworthy

Problem #1: No Power Calculation

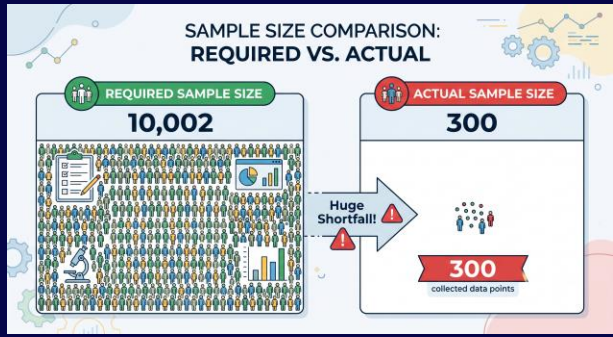
- The word “power” is not mentioned in the article, and it appears that no power calculations were done, as the number of users triggered is tiny
- See article Power Analysis is Essential at https://bit.ly/powerAnalysisEssentialEJW
- What is a “power calculation?
  - It is a simple formula that allows us to compute the minimum sample size needed
  - It requires several parameters, such as conversion rate, alpha (usually 0.05), and power (usually 80%), and one special parameter: the MDE: Minimum Detectable Effect
  - How do we choose the MDE? We can learn from companies that ran a lot of experiments:
    - At Bing, the average treatment effect was rarely over 0.3%
    - At Airbnb search, successful experiments improved conversion by 0.3%
    - The Analytics Toolkit reported a median lift of 0.1% from 1,001 experiments

    You should pick an MDE below the average or median, as this is the MINIMUM
  - Picking 0.3% will yield sample sizes that most small to medium organizations don’t have
  - For our community replication project, Trustworthy A/B Patterns, we picked 2% in most cases
  - As a rule of thumb, I recommend that you never pick an MDE over 5% (unreasonably optimistic)

Sample Size Required

- Given the reported conversion rate for Control of 55.7%, plugging this into a power calculator (e.g., https://www.evanmiller.org/ab-testing/sample-size.html) with 5% relative MDE yields a minimum sample size of…

  10,002 for the two variants



- How many did this experiment have?

  300 users: highly underpowered

  (Note, misleadingly, the author stated 54,000 users per variant, but only a small percentage reached the page where the change was made, so the triggered population, which is the sample size needed, was tiny.)

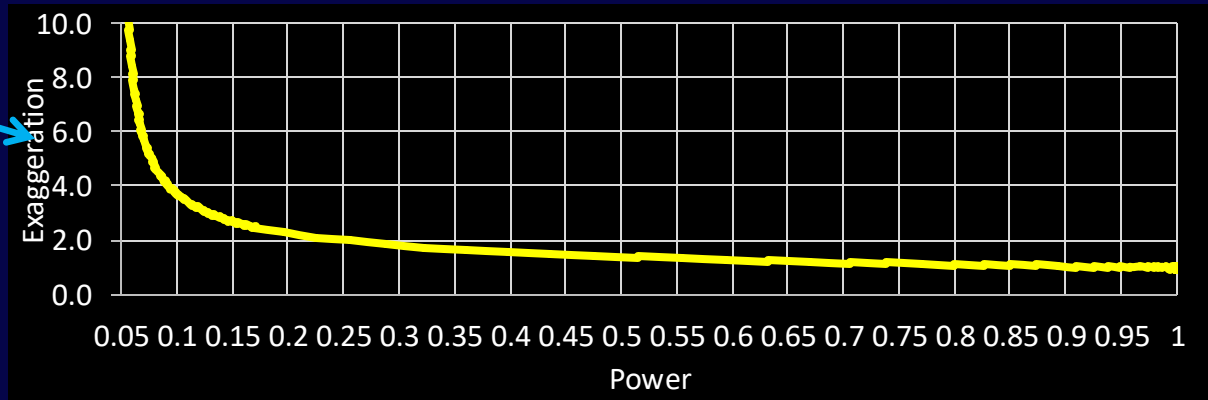
- We could also ask the reverse question: what power did this experiment have?

  6.9% power for the right tail, dramatically below 80% (see https://bit.ly/powerFromNCrazyEgg)

Two Problems with Low Power, Say 10%

1. You are unlikely to get stat-sig result.
   If you have 10% power, and there is a real effect of MDE size, you will only detect it 10% of the time (remember, 5% are stat-sig when there is no difference: type-I error)
2. Winner's curse: If you get a statistically significant result with power < 50%, the treatment effect must be exaggerated (spreadsheet, fill cell B20 with power)
   (Gelman and Carlin 2014, Andrews, Kitagawa, McCloskey 2023)

At 10% power, the Exaggeration Is ~4 times!

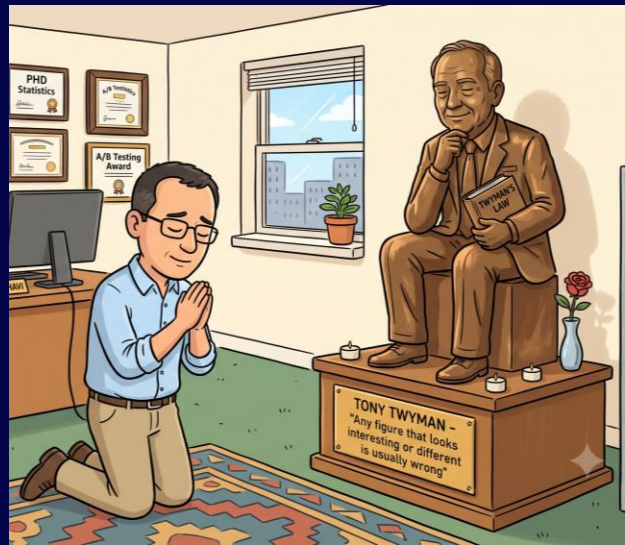


Expected Exaggeration in our Example?

- For our example
  - The EXPECTED exaggeration (type M error) is **5.0x**
  - The minimum exaggeration is 4.1x
  - The probability of a sign-error (you thought Treatment was better, but Control is better) is **9.6%**, vastly higher than the 1% implied by “99% confidence.”
- The problem: many A/B tests are being run, so with an alpha threshold of 0.05, we will see a lot of false positives that will exaggerate and make headlines

Praying to Twyman

- Documented success example at https://bit.ly/prayToTwyman
  - Yesterday I had 100% increase in conversion rate!
  - I flipped a coin 10 times and got 3 heads
  - I then prayed to my god called Twyman, flipped another 10 times, and got 6 heads
  - That's 100% increase!
  - Send me $1,000 (PayPal, Venmo accepted), and I will send you the prayer, and you'll benefit from amazing ROI! (10% of proceeds will be donated to the Education for the Gullible).
- I hope you realize the absurdity here
- But this is what many published results look like:
  - They don't do power calculations
  - They don't show you the 20 prayers they tried before finding the "correct prayer."



We Need Large Samples

Here is a table with the total sample size needed (two variants) given the conversion rate and 5% relative MDE

Conversion/ Success	Total users for A/B Test with 80% power, 5% relative lift	Example
50%	12,800	Initiating checkout to completing checkout
20%	51,200	Add-to-cart after visiting product detail page
10%	115,200	Initiating checkout
5%	243,200	Conversion to sale for a cheap product and tuned site
2%	627,200	Conversion to sale for $50+ product and tuned site
1%	1,267,200	Conversion to sale for poor site

Problem #2: Misinterpreting P-values

- The author claimed 99% confidence, which they interpreted as “we can be 99% certain the result reflects a genuine difference between the two pages, not a statistical accident.”
- This is wrong!
- A 99% confidence claim usually corresponds to a p-value of 0.01, but the p-value is a conditional probability that assumes the null hypothesis is true
- The probability described is FPR: the False Positive Risk
- It can be computed, but it needs another parameter: the prior probability of success
- We can ask: what is that prior probability at large organization?

FPR – False Positive Risk

Company / Source	Success Rate	FPR (alpha=0.05)	FPR (alpha=0.10) Optimizely's default
Microsoft	33%	5.9%	11.1%
Avinash Kaushik	20%	11.1%	20.0%
Bing	15%	15.0%	26.2%
Booking.com, Google Ads, Netflix, Optimizely (127K)	10%	22.0%	36.0%
Airbnb Search	8%	26.4%	41.8%

- The Median success rate in this table is 10% (there are really 4 entries there)
- When using alpha=0.05 for the median success, a statistically significant result is 22% likely to be a false positive
- The success rate is unknown but can be estimated.
  See https://bit.ly/FalsePositiveInABTests
- For our example with low power, the false-positive-risk is **87%** (compare to 1% claimed)

Problem #3: SRM – Sample Ratio Mismatch

- In an experiment with equal percentages assigned to Control/Treatment, you should have approximately the same number of users in each
- Real example:
  - Control: 821,588 users, Treatment: 815,482 users
  - Ratio: 50.2% (should have been 50%)
  - Should I be worried?
- Absolutely
  - The p-value is 0.0000018, so the probability of this split (or more extreme) happening by chance is less than 1 in 500,000 (the Null hypothesis is true by design)
  - See https://bit.ly/srmCheck for Excel spreadsheet to compute the p-value
  - ~6% of Microsoft experiments have an SRM, so this is a big problem (alpha threshold of 0.001). Paper describes how to diagnose SRMs
  - Convert.com added SRM check 4/7/2022: 6.5% of experiments have SRMs

Sample Ratio Mismatch in our Example

- In our example, which stated a 50/50 design, the number of triggered users was 176 and 124
- This is imbalanced: the SRM p-value is just below 0.003, so it is more statistically significant than the result itself, a dark yellow flag

How Bad is a Small SRM?

- Real example from Bing
  - Experiment ran and showed amazingly strong results for the top key metrics

Delta	p-value
+0.54%	0.0094
+0.20%	7e-11
+0.49%	2e-10
-0.46%	4e-5
+0.24%	0.0001

  - The ratio of users was 0.497 instead of 0.5
  - Close? Yes, but with almost a million users, the probability was 0.00002. Unlikely
  - How off could the numbers be from such a small imbalance?
  - The reason was a bot, so we excluded it and got a balanced set of users
  - Nothing was stat-sig. Imbalances are usually caused by very skewed “users”

Delta	p-value
+0.19%	0.3754
+0.04%	0.1671
+0.13%	0.0727
-0.12%	0.2877
+0.01%	0.8275

Problem #4: Twyman's Law

- In the book *Exploring Data: An Introduction to Data Analysis for Social Scientists*, the authors wrote that Twyman's law is

  "perhaps the most important single law in the whole of data analysis."

Any figure that looks interesting or different is usually wrong

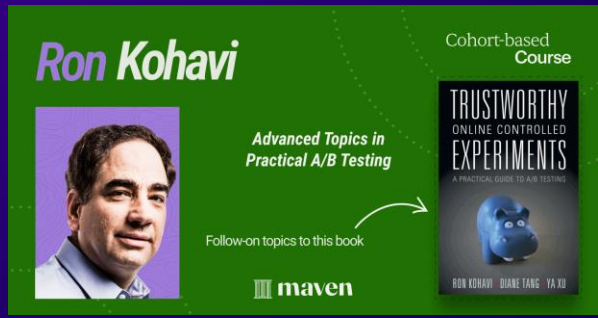
- A reported 44% lift on a solid baseline control conversion rate of ~56% should immediately trigger skepticism
- I have never seen a trustworthy experiment produce a true 44% improvement on an important metric unless it was fixing a broken experience

Other Checks

- Run A/A tests: split the users into two groups as in a regular A/B test but make B identical to A (hence the name A/A test)
  - If the system is operating correctly, then in about 5% of the trials, a given metric should be statistically significant with p-value less than 0.05
  - For statisticians: when conducting t-tests to compute p-values, the distribution of p-values from repeated trials should be close to a uniform distribution
  - Chapter 19 in my book is on A/A tests and learnings.
    Another good ref by Simon Jackson
- Bot traffic. At Bing, 50% of traffic in the US was bot generated, and 90% of the traffic in Russia and China was bot generated
- Novelty / Primacy effects – if the treatment effect is decreasing/increasing with time, it could be due to these effects.
  Reality: usually trends are just noise and not due to these

Want to learn more about A/B testing?

- I teach two online courses on Maven
- We will raffle three tickets to the Intro class during the Q&A
- Accelerating Innovation with A/B Testing ★★★★★ 4.8 (165)
  - $1,999: five sessions x 2.5 hours each
  - $250 discount for first 5 people to register with this URL: https://bit.ly/ABCourseRKGB
- Advanced Topics in A/B Testing ★★★★★ 5.0 (71)
  - $1,199: three sessions x 2 hours each
  - $200 discount for first 5 people to register with this URL: https://bit.ly/AdvABRKGB



Introduction to Luke

Luke Sonnet

GrowthBook's Head of Experimentation

Leading the team building experimentation tooling (statistics, queries, metrics)

PhD in social sciences, mostly causal inference

Previously on platform teams

Twitter
Experimentation Data Science

Meta
Survey weighting and ML ground truth validation

Trust is built on transparency and reliability

A trustworthy experiment gives you reliable information. Statistical significance is one signal we rely on, and you **control the threshold** (in many ways!)

0.05

THE STANDARD ALPHA

With **no true impact, 5%** of experiments will falsely appear significant

Should every experiment use 0.05?

If a feature is costly, pick a stricter threshold

Costly = ongoing support, dictates the product roadmap

Is your significance threshold reliable?

Do you regularly catch yourself saying...

> *Let's run it longer to see if it becomes significant.*

> *The goal metric is flat, but these leading indicators are positive, so let's ship!*

Problem 1: Peeking

Not sticking to your planned experiment sample size or runtime **inflates false positive rates**

TWO FORMS OF PEEKING

Stopping early

Looking early and stopping when you see stat-sig results

Running longer

Hitting the planned runtime, then deciding to ‘run it longer’

How big of a problem is it?

Problem 1: Peeking, how bad is it?

Let's imagine you run an experiment **with no impact** for your planned duration, but then say “let’s monitor to see if it becomes stat-sig?” and then check continuously.

Duration	Type I Error Rate
Planned duration	5%
Up to 2x planned duration	17%
Up to 5x planned duration	38%

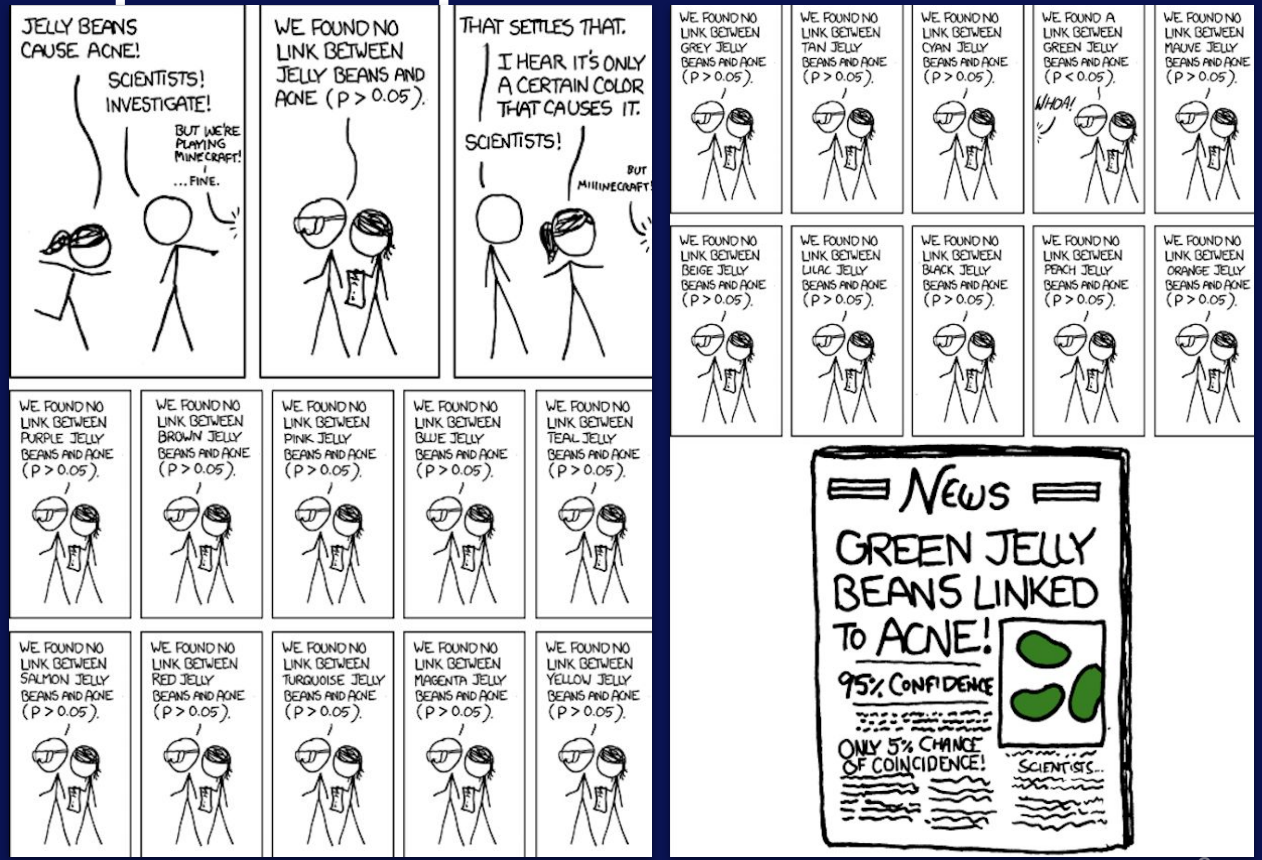
Netflix simulation showing 70% erroneous early stopping

Booking showing that minimal peeking can inflate estimated effects by 2.5x standard errors

Problem 2: Multiple comparisons

Test enough metrics and one will look significant by pure chance.

20 colors tested, one comes up "significant."



xkcd.com/882

Problem 2: Multiple Comparisons, how bad is it?

If you end an experiment if any metric is stat-sig, but actually in this experiment there is **no impact**, how often did you make the wrong call?

Number of Decision Metrics	Type I Error Rate
1	5%
2	10%
3	14%

The problems compound

Peeking and
Multiple Comparisons
actually compound!

Type I Error under previous slides' settings				
		Duration		
		1x Planned	Up to 2x Planned	Up to 5x Planned
Number of Decision Metrics	1	5%	17%	38%
	2	10%	30%	62%
	3	14%	42%	76%

The problems compound

Peeking and **Multiple Comparisons** actually compound!

Problem looks even worse when you consider the **False Positive Risk.**

FPR under previous slides' settings				
		Duration		
		1x Planned	Up to 2x Planned	Up to 5x Planned
Number of Decision Metrics	1	13%	32%	52%
	2	23%	47%	66%
	3	31%	57%	73%

Assuming 80% power, 15% win rate, only ship positive tests

Is your significance threshold reliable?

Do you regularly catch yourself saying...

> *Let's run it longer to see if it becomes significant.*

> *The goal metric is flat, but these leading indicators are positive, so let's ship!*

Solutions?

Do power analysis and stick to the plan

Strive for a single goal metric and design power for that goal metric

Multiple goals? Consider multiple comparisons corrections

Want to peek and stop early?
Use alpha-spending or sequential testing depending on your use case

Solutions?

1. Be honest: what are the costs of my experiment?
2. Do those costs imply stricter or more relaxed thresholds?
3. Does my process **honor** those significance thresholds?

Raffle Winners & Questions

Drop your questions in the Q&A panel, we'll answer as many as we can

TODAY'S SPEAKERS

Ronny Kohavi
VP & Technical Fellow Airbnb,
Microsoft, & Maven Instructor

Luke Sonnet
Head of Experimentation
GrowthBook

NEXT WEBINAR

How The Philadelphia Inquirer scaled experimentation across the organization

Thursday, August 20 · 9am PT / 12pm ET

Scan to register

Thanks for joining

The recording will be emailed to you in the next few days

STAY CONNECTED

linkedin.com/company/growthbook

growthbook.io/our-community

growthbook.io

GET STARTED

Start experimenting today for free!

Scan to start