The Joy of Proving Yourself Wrong
A Case Study in A/B Experimentation

@jennydove
#ProveWrong

Agile 2018
San Diego
#Agile2018
Maximize Your Time Here

- Participation is expected
- Please sit near the front
- Grab a pen and some paper
- Give others a turn
- Form a group of 4-5 people for your team
Good Experimentation Helps You Close the Feedback Loop
People say A/B Testing is Easy

I haven't had a test go that simply, pretty much ever.
Take pride in measuring correctly, not in being correct
Case Study
Which do we need to test?

Adding Apple Pay to the App
We want to add Apple Pay to the app to simplify payment. Right now we only accept credit cards.

Visual style update on checkout
Updating the checkout page to match our new brand. This includes typeface, spacing, and color.

Asking for license plate after purchase
Some spots require a license plate to use. Today it’s required before payment, but want to move it.

New error messages for cards
All payment errors now return “Server Error.” Update to say “Card Declined” and “Server Timeout.”
License Plate Entry

Don't Get Towed
To validate your reservation, you must enter your license plate information before you park.

License Plate

Add Promo Code

Payment Method
VISA*4242

Amount:
$8

I Don't Know

Book for $8.00

By purchasing, you agree to SpotHero's Terms & Conditions.

Reservations

Current Reservations

718 W Monroe St. - Lot
May 6, 13:30 - 23:59

We've updated your times for free!

Oversized fee charged at location

Book with Apple Pay

Edit
Cancel
Conversion Funnels
Group Exercise: Map the conversion funnel

- Map out the conversion funnel on a piece of paper
- Determine what the relevant metrics are to advance to the next step of the funnel
- Are there other metrics we should track that aren’t tied to a funnel step in particular?
Map the conversion funnel

- Ran a search
- Selected a spot
- Completed Checkout

- Called Support
- Got Towed
Quick Reference

- Today’s conversion rate
  - 9.0%
- App-Wide Traffic
  - 5k users per day
- License Plate Required Lots
  - 20% of purchases
  - 30% of spots
What is the hypothesis?
Elements of a good hypothesis

- What we are measuring
- Why we are measuring it
- What we expect the impact to be
- How long we think it will take
Test Setup Checklist

- When is the test supposed to be triggered?
- What counts as a conversion?
- What percentage of the population should see the test?
- What is the distribution of people in versions A and B?
- What level of precision do we need?
- What all could go wrong?
375 basis point improvement

After running the test for 2 days we see a dramatic improvement in conversion!
We double-counted conversions
Why would double-counting conversions cause the test results to appear skewed?
How long will the test take?
Hand out a few rulers
Have people measure the dimensions of their wallets/phones/something
Ask for the most precise number they can come up with
See if anyone provides a CI
Push them to be more and more precise, emphasize precise, not exact.

Things I want to explain:

1. How making sure the test measures and triggers at the right time matters a ton
2. That by defining your expectations ahead of time, you can have a better sense of whether your data is truthful or not
3. 
More on confidence intervals

Hand out a few rulers

Have people measure the dimensions of their wallets/phones/something

Ask for the most precise number they can come up with

See if anyone provides a CI

Push them to be more and more precise, emphasize precise, not exact.
Small Group Discussion: How precise should we be?

What are the costs of being more precise? What do we give up by taking longer?

How do we decide what is good enough? What’s the best process around that?

What percentage of the population should get the test? How do we bucket users? What distribution and allocation?

Use a tool to guess how long the test will run based on your precision.
Group discussion:

How do you know if your data can be trusted?

In this case, we knew the data wasn’t right.

When can we trust the data and test results and when do we need to go back and look again?

What are the clues that something might be wrong?
Data Accuracy is your Top Priority

- There’s no point in doing an experiment if you can’t measure the results

- Check the data:
  - Is the experiment triggered at the right time?
  - Are the results being measured correctly?
A Bug! What do we do?

When a customer updates their license plate after checkout, it updates on our end but doesn’t appear updated to them.
Fix and Relaunch

- This impacts what we’re measuring with regards to customer calls
- Therefore it interferes with our experiment
Exercise:
Given the bug, decide whether we need to restart the test or not

- Customers with more than 3 credit cards on file can’t check out in either version
- 10% of our users conversions aren’t showing up in the experiment version
- Customers with more than 1 saved license plate can’t switch it after checkout
- SpotHero staff member actions are also getting recorded in experiment results
A marketing campaign caused app opens to increase by 10% mid-test, but overall conversion went down 5%.

- External variables can have an impact on your test
- External variables include:

  - Promotions
  - Weather
  - News and PR
Small Group Discussion: When do you control for external variability and when are you okay?

In your groups, come up with guidelines that you can use to figure out when external variability should be considered test interference and when it can be ignored.
The Next Experiment

Do we need to wait for this experiment to end before we run the next one?

Adding Apple Pay to the App
We want to add Apple Pay to the app to simplify payment. Right now we only accept credit cards.

Visual style update on checkout
Updating the checkout page to match our new brand. This includes typeface, spacing, and color.

New error messages for cards
All payment errors now return “Server Error.” Update to say “Card Declined” and “Server Timeout.”
Multiple experiments can run simultaneously on the same population if they measure against different conversion events.

If they are for the same conversion event, you need to divide your population. This means the test will take longer.
Results are in!

<table>
<thead>
<tr>
<th></th>
<th>Conversion</th>
<th>Tows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Purchase License Plate</td>
<td>9.21%</td>
<td>4</td>
</tr>
<tr>
<td>Post-Purchase License Plate</td>
<td>10.65%</td>
<td>6</td>
</tr>
</tbody>
</table>

N = 15,764
The Inconclusive Test
Conclusion
Group Discussion - Lessons Learned

- Incorrect data collection
- External variability
- Careful test setup
- Statistical significance
- Bugs in your product
- Running multiple experiments
Just because you have results doesn’t mean you’re done

Check your data

Plan your next iteration

Share your results
The importance of execution

1. Data must be accurate or you are wasting your time
2. Isolate your variables and make sure you know what you’re testing
3. Know when you have confounding variables versus noise
4. There are going to be hiccups— it’s all about how you handle them.
Appendix
More on A/B Experimentation

- **A/B Testing at Duolingo**
- **A/B Sensei** - Detailed information on the statistics and process behind A/B experimentation
- **Designing with Data** - Great book going through experiment design in more detail