Experimenting on Users

Zijian Ding and Shiwei Zhou

Learning Goals

- Importance of controlled experiments on the web
- A/B testing
- MultiVariable testing
- Practical guide to conducting online experiments
- “Robustness” test and “Viaversary” test

Controlled experiments on the web: survey and practical guide

Kohavi, Longbotham, Sommerfield, Henne

To treat or not to treat?

To treat or not to treat on the web?

What they said...

“The difference between theory and practice is larger in practice than the difference between theory and practice in theory.”

– Jan L.A. van de Snepscheut

What they said...

“Almost any question can be answered cheaply, quickly and finally, by a test campaign. And that’s the way to answer them – not by arguments around a table. Go to the court of last resort – buyers of your products.”

– Claude Hopkins, Scientific Advertising, 1923

What you said...

“Generally, companies that support these kinds of experiments (controlled experiments) are more in touch with their user's needs and understand better how their users interact with their platforms."

– Fiona Cisternas

To treat or not to treat on the web?

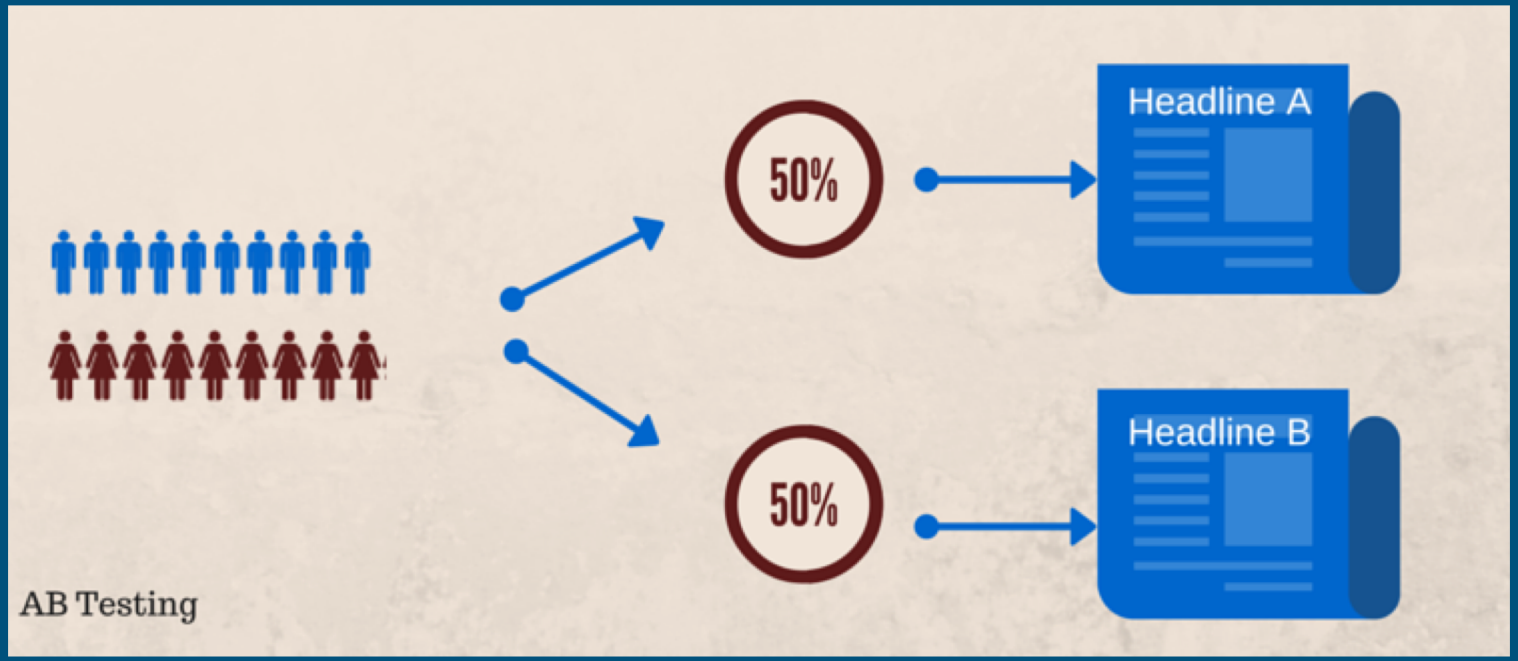
“Sufficient”:

- The web provides an unprecedented opportunity to evaluate ideas quickly using controlled experiments

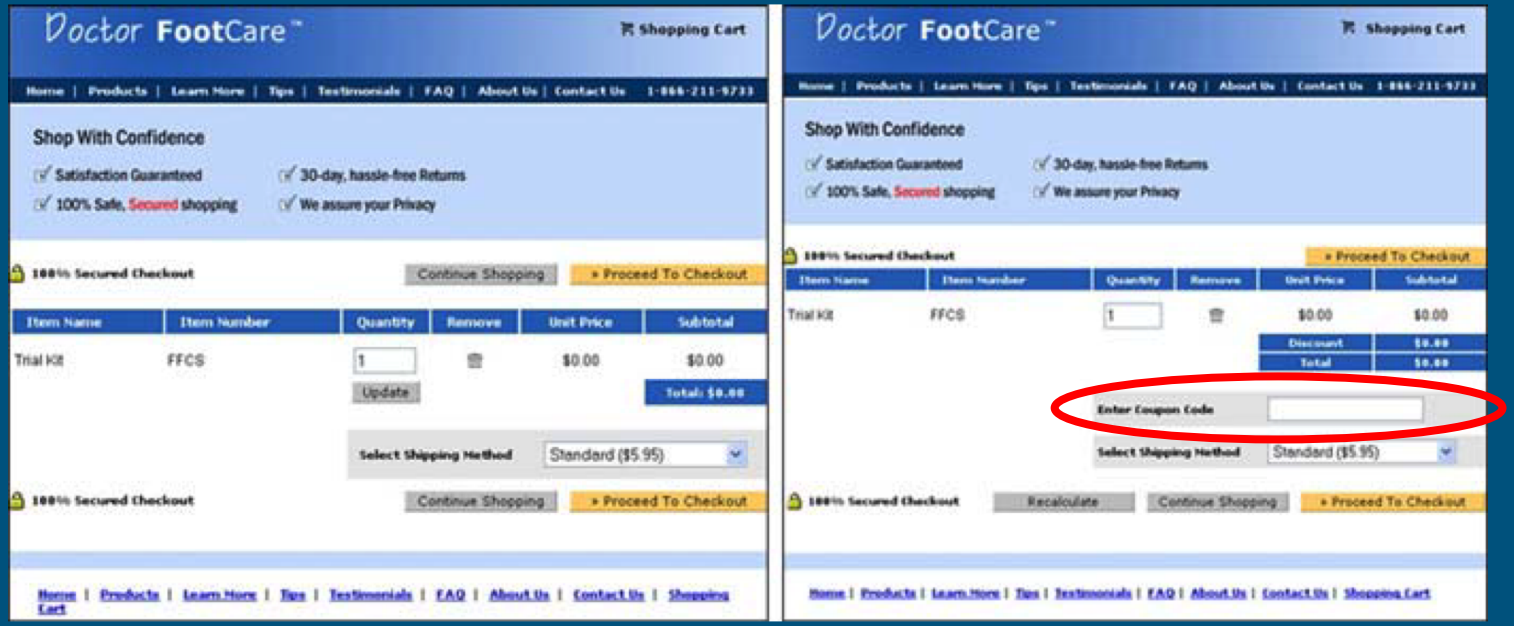
“Necessary”:

- Difficult to predict user behaviors on the web
- Difficult to make the tradeoff (e.g. MSN homepage ads)

A/B testing: simplest controlled experiment



A/B testing: simplest controlled experiment



lost 90% revenue

Discussion: which one creates more revenue

40% better

Why no incentive is better?

Many people who play a popular game like SimCity don’t play any other games.

The 20-percent-off offer didn’t resonate with them.

https://blog.hubspot.com/marketing/a-b-testing-experiments-examples

Do you really know A/B testing?

Techniques of A/B testing

- Run A/A tests (null tests)
- Minimize sample size
- Treatment ramp-up
- Auto abort
- ...

Run A/A tests

"Two groups, one experience"

- Assess the variability for power calculations
- Test the experimentation system

Minimize sample size

- Use Overall Evaluation Criterion (OEC) components with inherently lower variability
- Lower the variability of the OEC by filtering out users
- Set Treatment and Control groups to be equal size

Determine the minimum sample size

the variance of the OEC

n is the number of users in each variant

$$n = \frac{16\sigma^2}{\Delta^2}$$

sensitivity: the amount of change you want to detect

confidence level: 95%
power: 80%

Determine the minimum sample size - Example

- E-commerce site, 5% of users who visit during the experiment period end up purchasing
- Those purchasing spend about $75
- Standard deviation is $30
- Confidence level: 95% & Power: 80%
- How many users needed to detect a 5% change to revenue?

$$n = \frac{16\sigma^2}{\Delta^2}$$

Determine the minimum sample size - Example

409,000

$$n = \frac{16\sigma^2}{\Delta^2}$$

16*30^2

(75*0.05*0.05)^2

Treatment ramp-up

- Initiated with a small percentage of users assigned to the Treatment(s)
- Increased that percentage gradually
- Analyze the data at each step
- Make sure there are no severe problems with the Treatment before exposing it to more users

Auto abort

- Reduce the percentage of users assigned to the underperforming Treatment to zero
- Reduce risk of exposing more users to some error

“Fail often, fail fast."

- Don Norman

Techniques of A/B testing

Q: How to confirm the absence of fleet-related effects?

A: Run A/A tests (null tests)

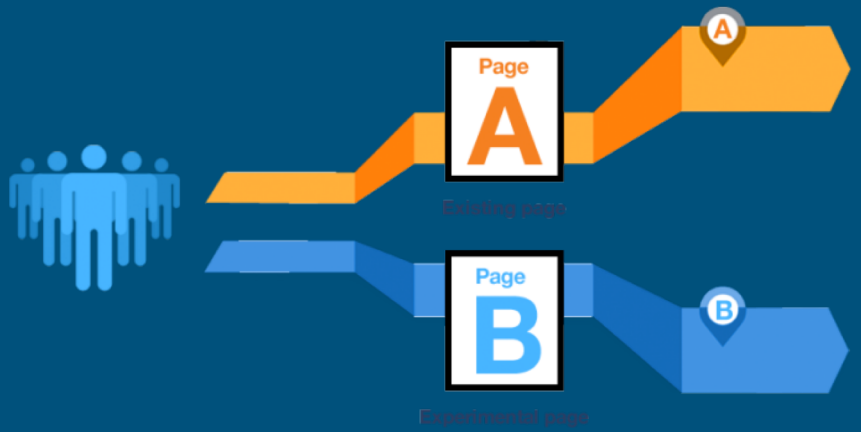
Q: If you run an A/B test and apply treatment ramp-up, when there is no difference at 10%/90% (treatment/control), what should you do?

A: Keep ramp-up until 50%/50% if no severe bugs are observed in the process

Discussion

Turn to your neighbors and spend 60s discussing following question:

Will you apply A/B testing to your research project? If so, how to apply it and how many participants do you need in your project?

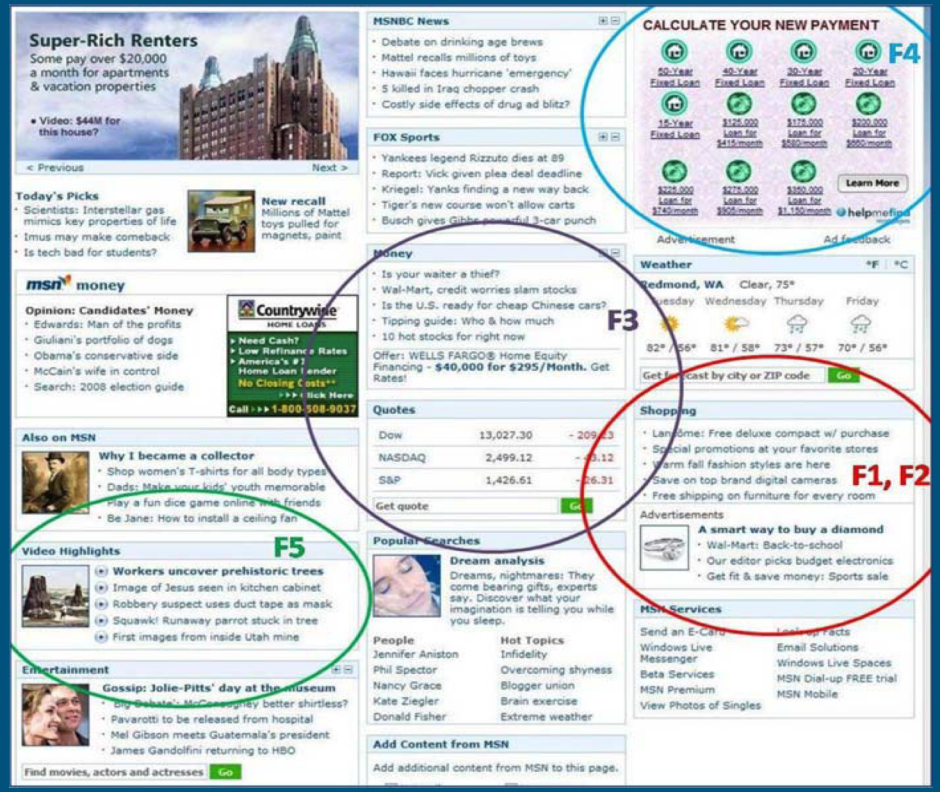


MultiVariable Testing

VS

MultiVariate Testing

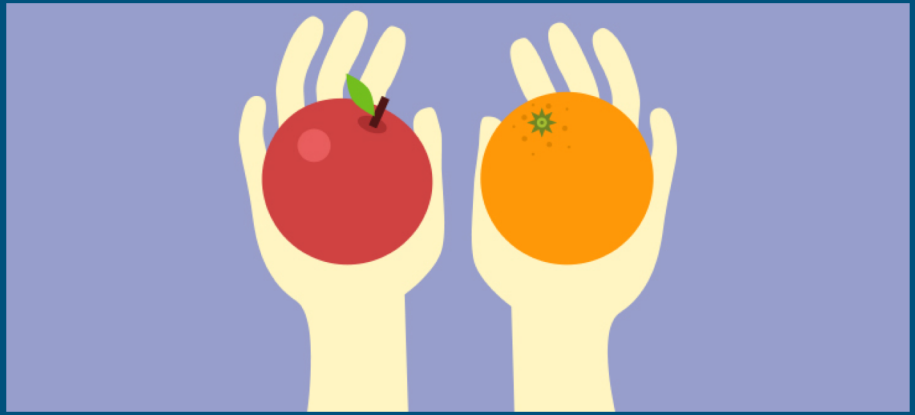
MultiVariable Testing



Discussion

Turn to your neighbors and spend 30s discussing following question:

What are the pros and cons of MultiVariable testing compared with A/B testing?



MultiVariable Testing

Pros:

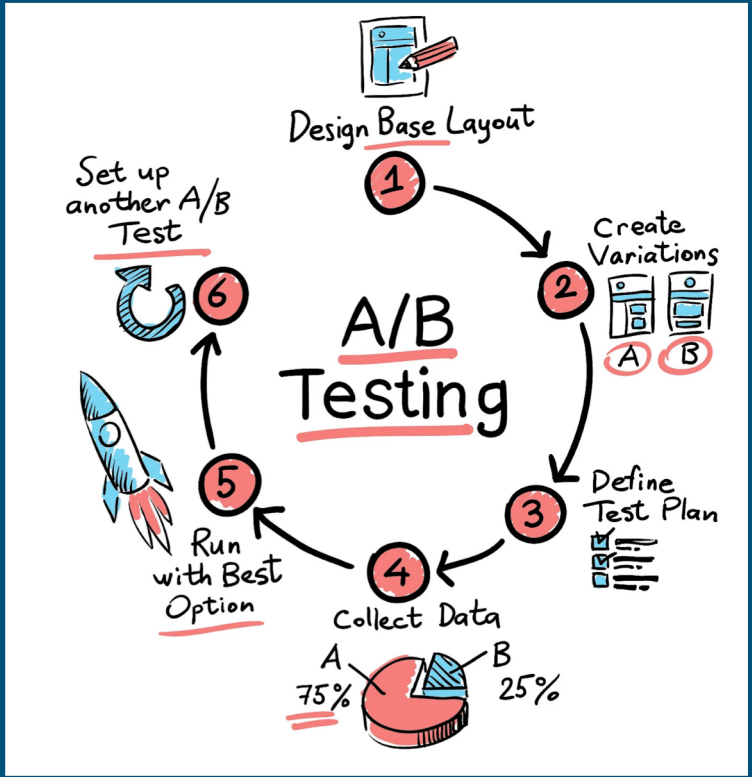
- Test many factors in a short period of time
- Estimate interactions between factors

Cons:

- Some combinations of factors may give a poor UX
- Analysis and interpretation are more difficult
- It can take longer to begin the test

Practical guide to conducting online experiments

- Randomization algorithm
- Assignment method
- Data path



Frustration-based Promotions: Field Experiments in Ride-Sharing

Maxime C. Cohen, Michael D. Fiszer
Baek Jung Kim

The set-up

- Users experienced frustration in ride-sharing
  - Waiting too long for the ride
  - Spending too long in the ride
- How to keep users using your app?
  - Do nothing
  - Communication
  - Credit
  - Waived
  - Discount

ETA(Estimated Time of Arrival) error=

actual waiting time - expected waiting time

VGR (Via Google ratio)=

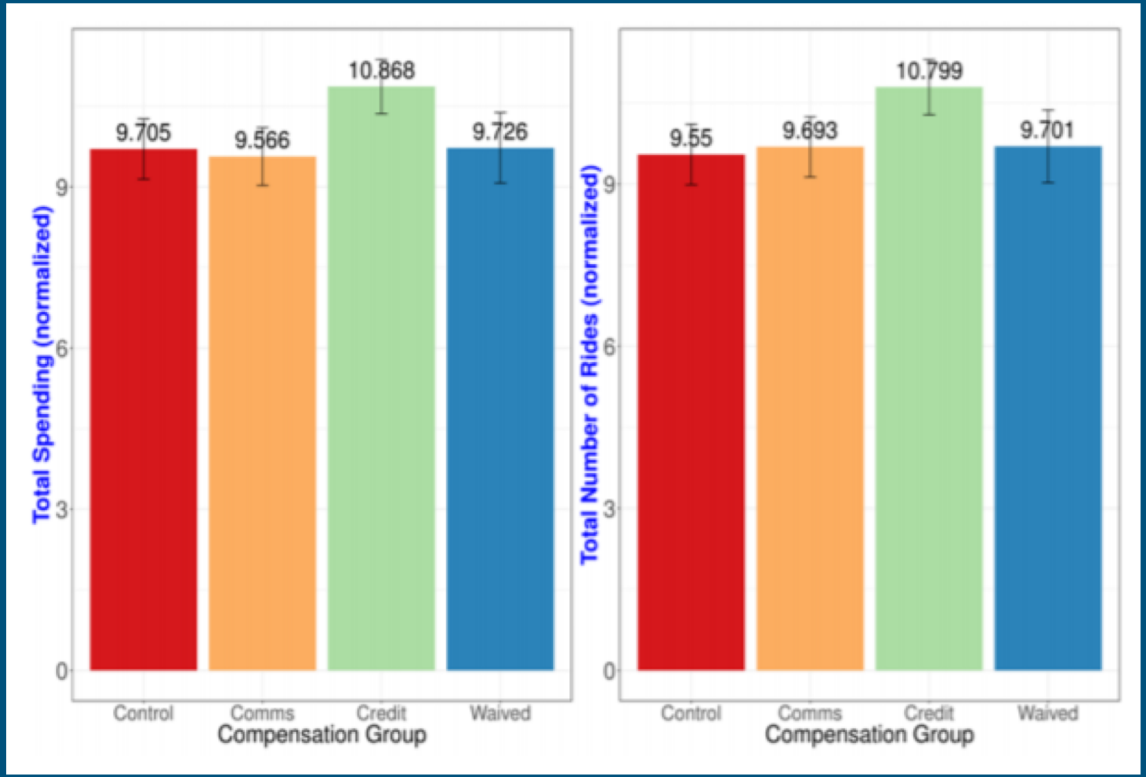
actual riding time / google map estimation

Experiment 1

New York City

- Control (961), Comms (992), Credit (1,341), and Waived (649)
- ETA=10mins
- VGR=2
- Credit = 5$

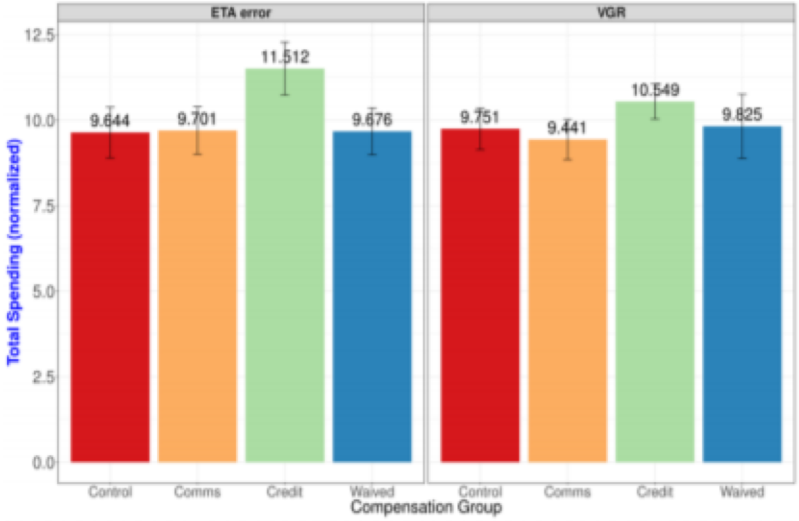
Credit compensation wins

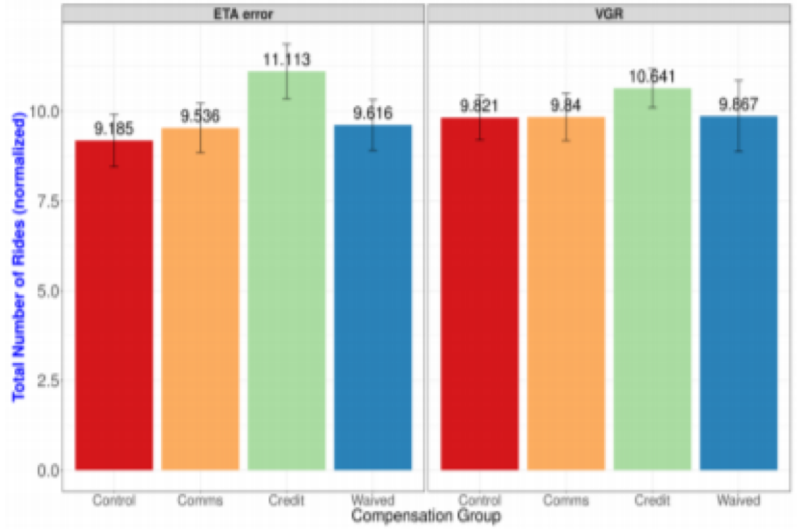


User bias can be influential

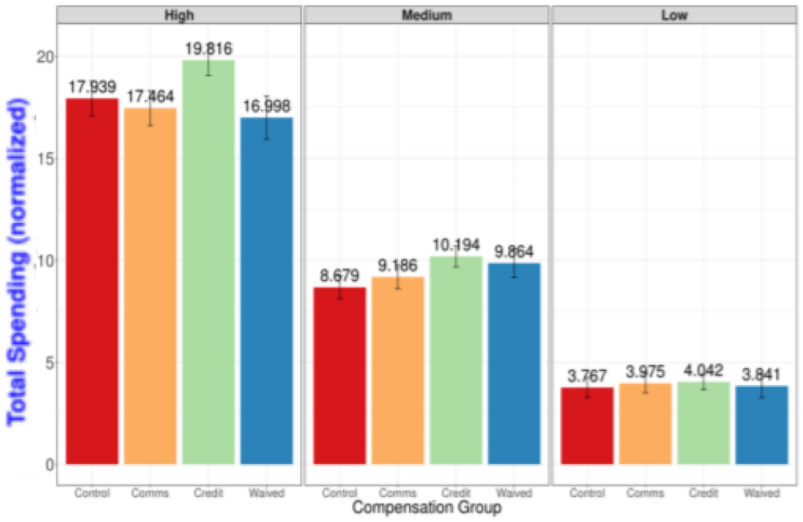
Result is more distinctive in ETS error

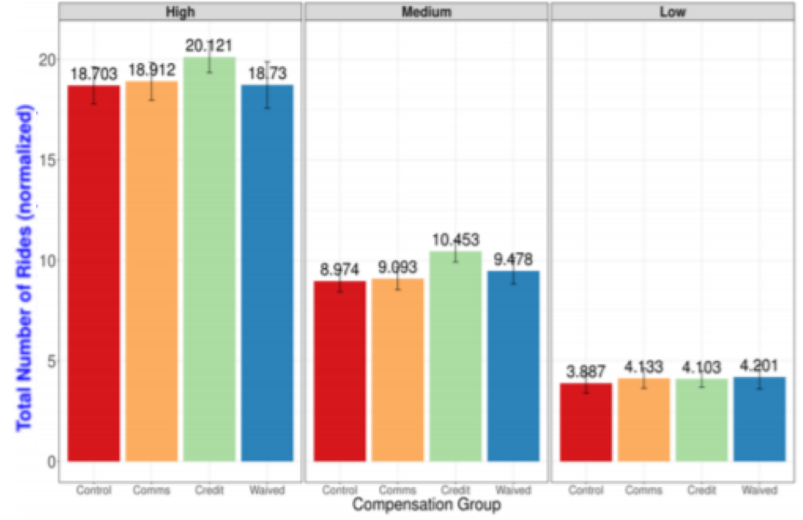
- Users are more likely to blame driver for long waiting, but not blame them for long ride
- Should we eliminate bias?



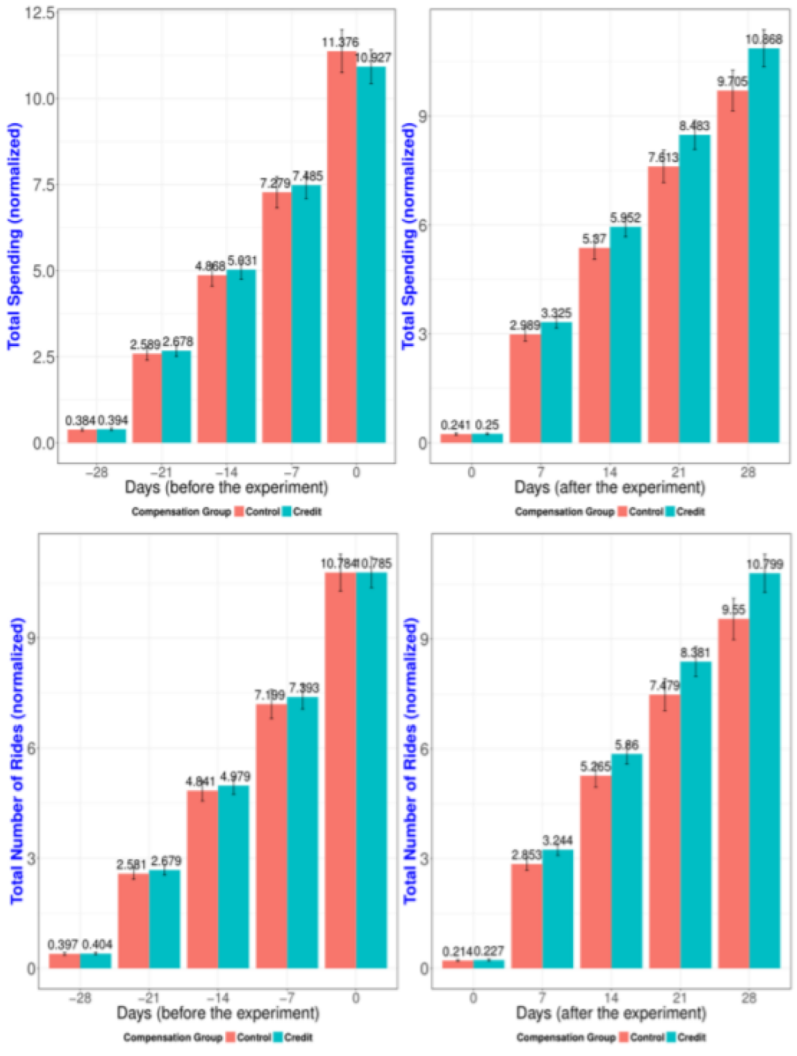


Most effective on high frequency riders

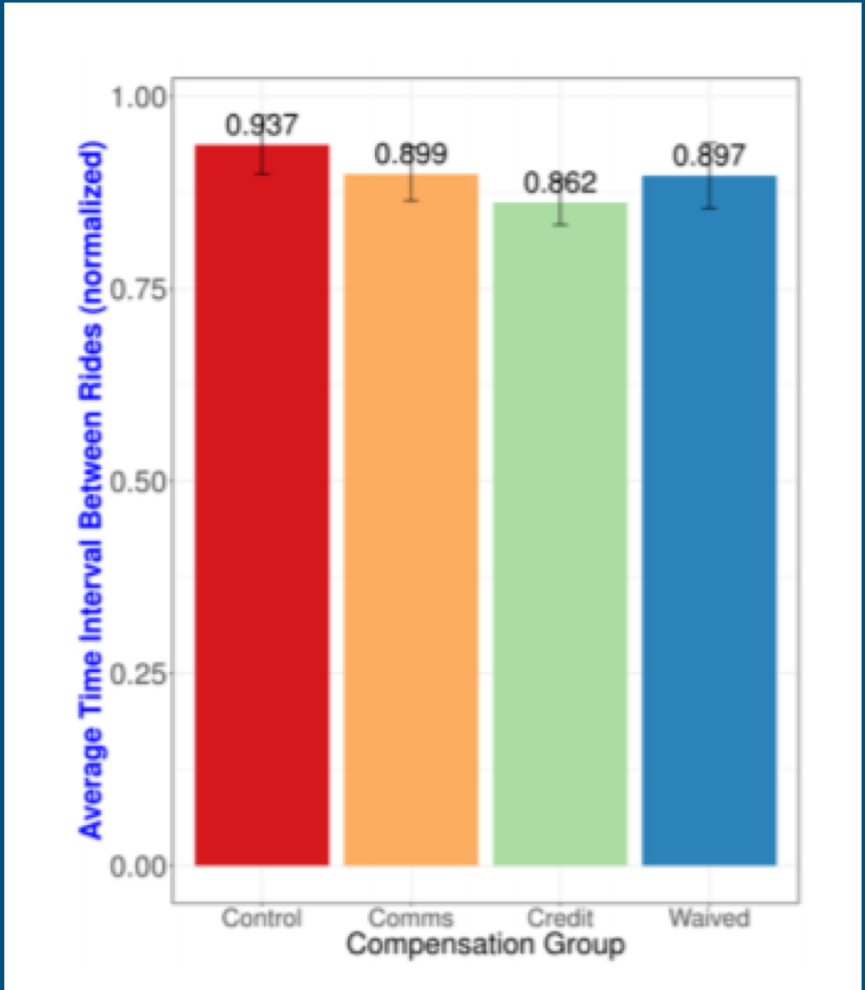




More effective in later days



Promote higher frequency of use (shorter interval between uses)



Features of credit compensation

- Effective in frequent riders
- Promote higher total number of rides
- Promote higher total spending
- Promote higher using frequency

Using math to find weight for each variable

- OLS regression

$$y_i = \alpha + \beta_1 \text{Comms}_i + \beta_2 \text{Credit}_i + \beta_3 \text{Waived}_i +$$
$$+ \gamma_1 \text{Pre-experiment-Rides}_i + \gamma_2 \text{Pre-experiment-Rides}_i^2 + \gamma_3 \text{ETA-error}_i + \mu_i + \epsilon_i,$$

- difference-in-difference

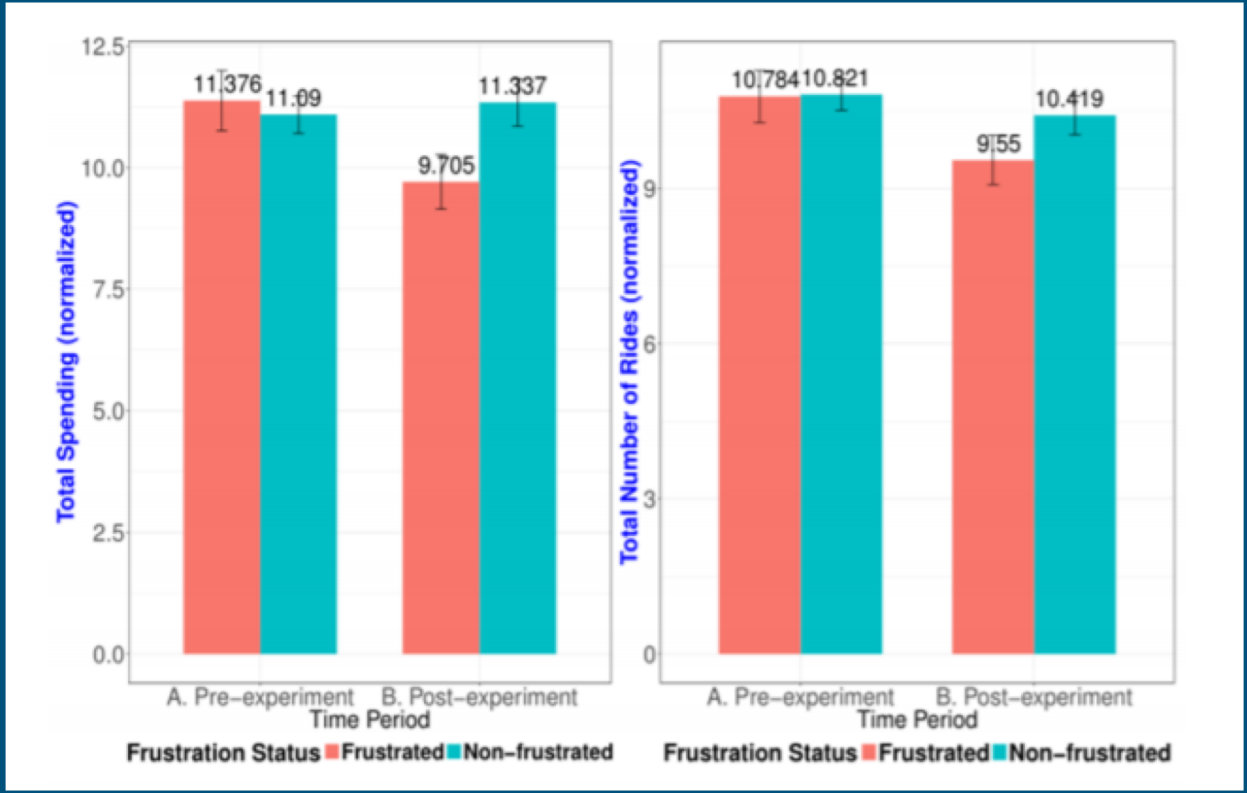
$$y_{it} = \alpha + \beta_1 \text{Comms}_i + \beta_2 \text{Credit}_i + \beta_3 \text{Waived}_i + \gamma \text{After-experiment}_t +$$
$$+ \delta_1 \text{Comms}_i \cdot \text{After-experiment}_t + \delta_2 \text{Credit}_i \cdot \text{After-experiment}_t +$$
$$+ \delta_3 \text{Waived}_i \cdot \text{After-experiment}_t + \epsilon_{it},$$

Revisit credit features again

Credit has largest coefficient of 1.2 in OLS regression

Has coefficient of 1.9 for ETS section and 0.6 for VGA section

Compensating non-frustrated is useless

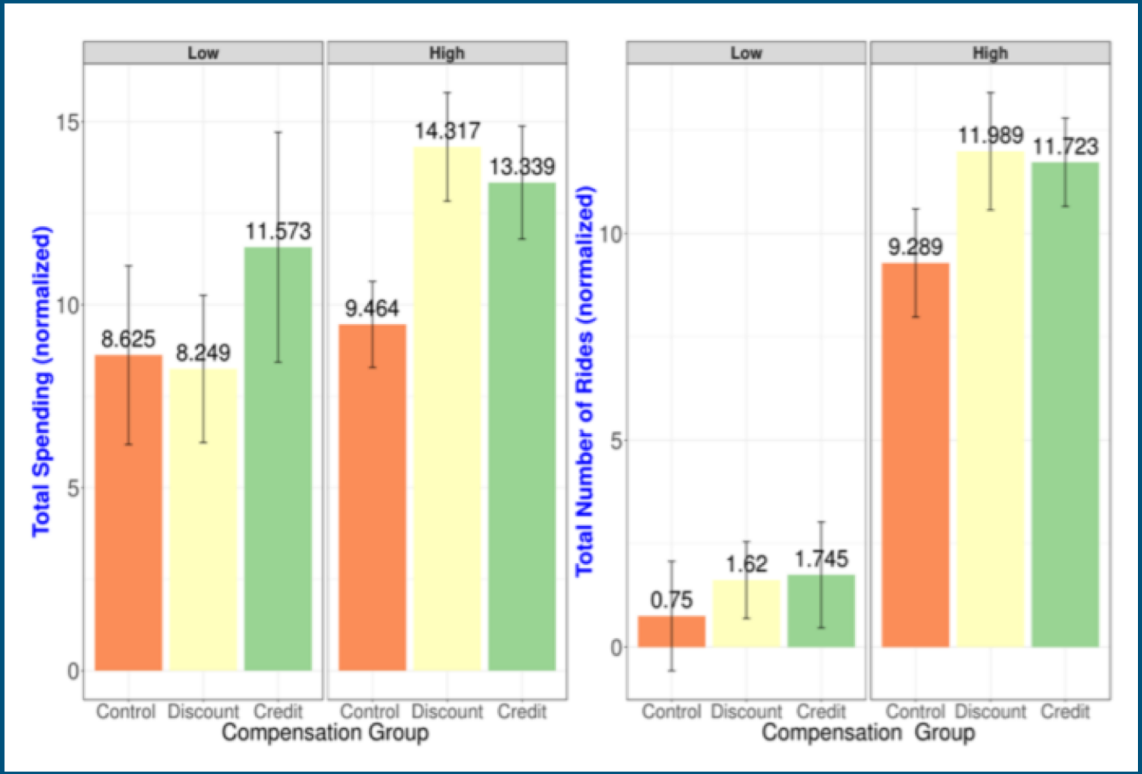


Experiment 2 (robustness test)

Washington D.C

- Control(308),Discount(342),Credit(298)
- ETA changed from 10min to 8min
- Add discount as new way for compensation
- Delete VGR section, commons and waives conditions

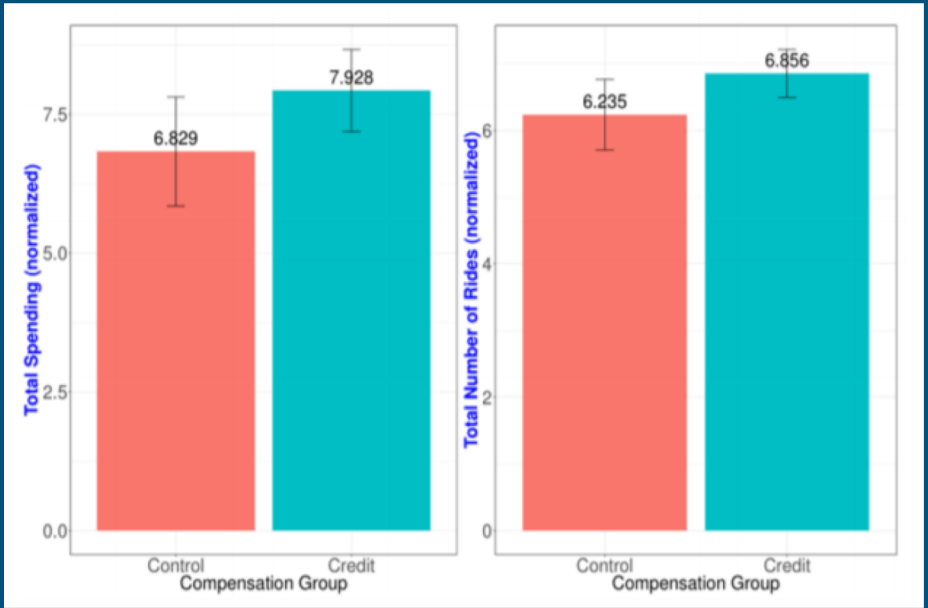
Is the result robust enough?



Experiment 3 (viaversary test)

Test “Viaversary" of the previous result

- Compensate random users
- Control (177) and Credit (428)



Discussion

Turn to your neighbors and spend 30s discussing following question:

Did Cohen follow rules in Kohavi's paper? Where he did and where he didn't? Whose method you agree more?

Is HCI the right path to a good design?

“I'll miss working with the incredibly smart and talented people I got to know there. But I won't miss a design philosophy that lives or dies strictly by the sword of data."

--Douglas Bowman (ex Google designer)


THANK
YOU