Web experiments

Benjamin Weggersen, Pallavi Agarwal

Learning goals

- Focus web experiments around organizational goals
- How to facilitate shorter testing cycles
- When to use A/B testing and when to use MVT
- The Facebook experiment and its controversy
- The debate on ethics
- Goodbye Google - data driven vs experience

Controlled experiments on the web

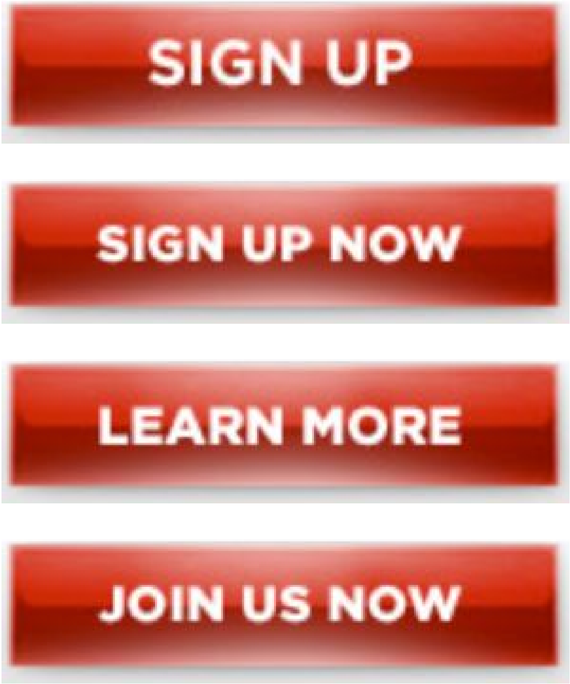
Ron Kohavi and others

...the ability to experiment easily is a critical factor for Web-based applications. The online world is never static. There is a constant flow of new users, new products and new technologies.

– Hal Varian, 2007


OBAMA'08
GET INVOLVED
Obama'08
JOIN THE MOVEMENT
Email Address
Zip Code
SIGN UP
PAID FOR BY OBAMA FOR AMERICA
CONTINUE to WEBSITE

Which button had the highest sign-up rate?



The results

Relevance Rating	Variation	Est. conv. rate	Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Button 5 / 5	Original	**7.51%** ± 0.2%	—	—	5851 / 77858
	Learn More	**8.91%** ± 0.2%	100%	18.6%	6927 / 77729
	Join Us Now	**7.62%** ± 0.2%	73.5%	1.37%	5915 / 77644
	Sign Up Now	**7.34%** ± 0.2%	13.7%	-2.38%	5660 / 77151

Media: "Get Involved"

Media: “Family”

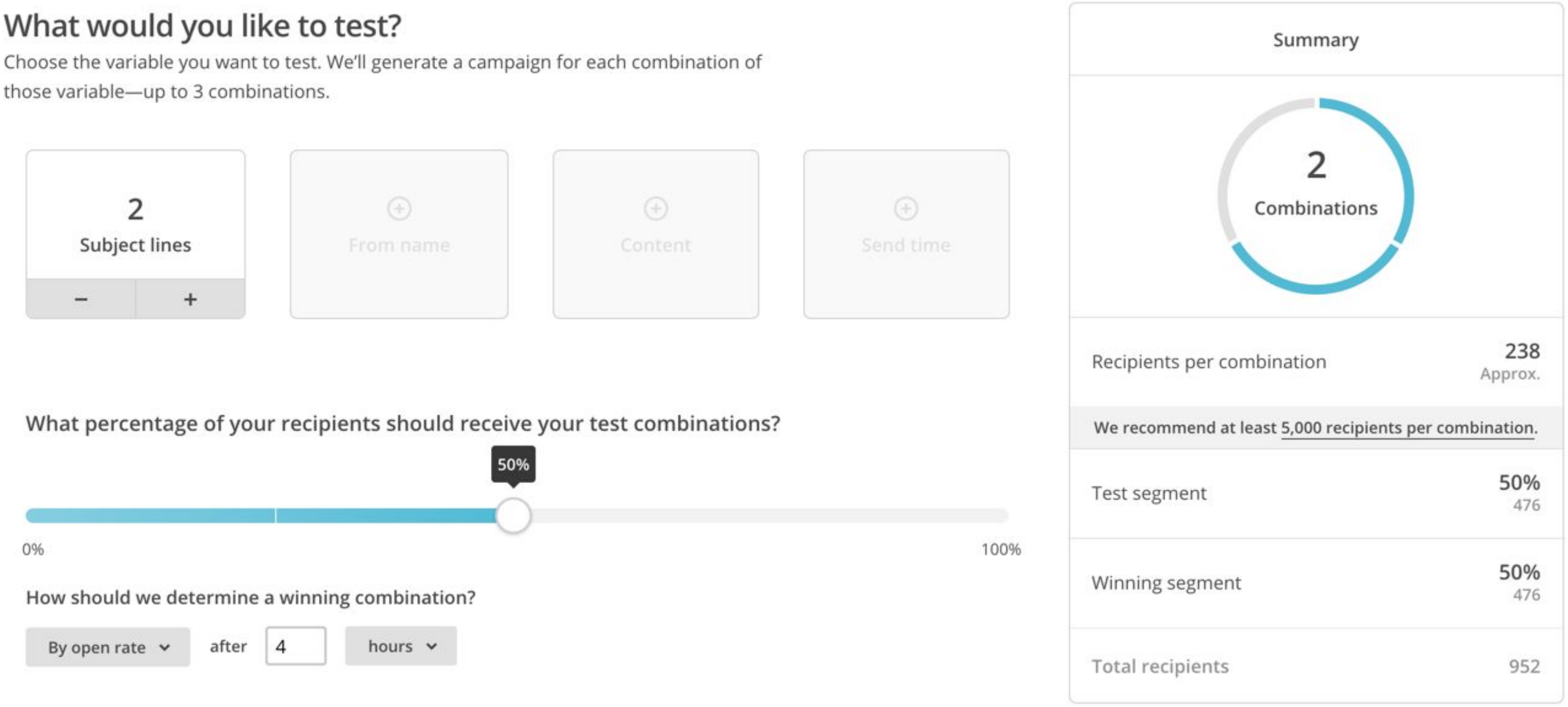
Media: “Change”

The results

Relevance Rating	Variation	Est. conv. rate	Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Media	Original	**8.54% ± 0.2%**	—	—	4425 / 51794
5 / 5	Family Image	**9.66% ± 0.2%**	100%	13.1%	4996 / 51696
	Change Image	**8.87% ± 0.2%**	92.2%	3.85%	4595 / 51790
	Barack's Video	**7.76% ± 0.2%**	0.04%	-9.14%	3992 / 51427
	Sam's Video	**6.29% ± 0.2%**	0.00%	-26.4%	3261 / 51864
	Springfield Video	**5.95% ± 0.2%**	0.00%	-30.3%	3084 / 51811


OBAMA'08
CHANGE
WE CAN BELIEVE IN
JOIN THE MOVEMENT
Email Address
Zip Code
LEARN MORE
PAID FOR BY OBAMA FOR AMERICA
CONTINUE to WEBSITE

Demo: Mailchimp



Section 1: Overall Evaluation Criterion

- A single metric
- Short-term vs. Long-term goals
- Choose components with lower variability
- Implications for organizations

Section 1: Overall Evaluation Criterion

A single metric

Page clicks	+	Conversion rate	+	Repeat visits	=	OEC
0.15		0.45		0.40		

Section 1: Overall Evaluation Criterion

Short-term vs. Long-term goals

- A good OEC should … include factors that predict long-term goals, such as predicted lifetime value and repeat visits.
- Example:
    - How might this influence ad revenue?
    - How might this influence repeat visits?

Section 1: Overall Evaluation Criterion

Choose components with lower variability

variance

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

sensitivity

Section 1: Overall Evaluation Criterion

Components with lower variability

Revenue

$$n = \frac{16 \times \$30^2}{(\$3.75 \times 0.05)^2} = 409{,}000$$

Conversion rate

$$n = \frac{16 \times \left(0.05 \times (1 - 0.05)\right)}{(0.05 \times 0.05)^2} = 122{,}000$$

Section 1: Overall Evaluation Criterion

Implications for organizations

- In formulating an OEC, an organization is **forced to weigh** the value of various inputs and decide their relative importance.
- This hard up-front work can **align the organization** and **clarify goals**.

Activity: Overall Evaluation Criterion

2 min

Break into groups of three. Each of you takes the role of either a CEO, a Marketing Director, or a Designer. Give weights to these criterions and argue why. You all work for Amazon.

- Page views
- Repeat visits
- Conversion rate (percentage of visits that include a purchase)
- Units purchased
- Revenue
- Bounce rate (percentage of users who exits after one page visit)

Section 1: Overall Evaluation Criterion

From the commentaries

- While it's clear when you're performing A/B tests you must have something measurable and thus comparable, blindly picking a "good enough" metric may not be the right answer. The key is achieving an overall improvement (with all stakeholders in mind; the company and the users).
  Vincent Chan

Section 2: Ramp up and auto-abort

- Gradual increase
- Real time analysis with auto-abort
- Requires good hash function
- Implications for organizations

Section 2: Ramp up and auto-abort

Gradual increase

99.9% / 0.1% ⇨ 99.5% / 0.5% ⇨ 97.5% / 2.5%

⇨ 90% / 10% ⇨ 50% / 50%

Section 2: Ramp up and auto-abort

Real time analysis with auto-abort

- At each step you can analyze the data to make sure there are no egregious problems with the Treatment before exposing it to more users.

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

sensitivity

Section 2: Ramp up and auto-abort

Real time analysis with auto-abort

Detect 1% change in OEC	Detect 20% change in OEC
1/20th of running time	1/400th of running time
~17 hrs	< 1 hr

Section 2: Ramp up and auto-abort

Requires good hash function

- Support monotonic ramp-up
- Slowly assign users to the Treatment
- New assignments should not change previous assignments

Section 2: Ramp up and auto-abort

Implications for organizations

- Allows organizations to make bold bets and innovate faster
- Auto-abort lets you to more confidently test on larger groups of users, thus reducing running time
- Integrate customer feedback directly in the development process through prototypes and experimentation

Activity: Ramp up and auto-abort

2 min

Break into groups of three. Ramp up and auto-abort allows you to iterate much faster, and still have statistical power. Are shorter tests always preferred? Why/why not?

Section 2: Ramp up and auto-abort

From the commentaries

- The reason why 50% is ultimately chosen as the fraction to ramp up to is suggested by the author to **maximize the power** of an experiment while simultaneously **minimizing the running time**.
  Many students wrote this

- … in product design and experimentation [it] is very important that we test and experiment with intention to fail quickly allowing ourselves to adjust and change accepting / rejecting ideas.
  Irfan Mulic

Section 3: A/B test or MVT

- How are they different?
- Interaction between factors
- Bold bets and very different design

Section 3: A/B test or MVT

How are they different?

Button A	Button A	Button B	Button B
Image A	Image B	Image A	Image B

Section 3: A/B test or MVT

Interaction between factors

- Two factors interact if their combined effect is different from the sum of the two individual effects.
- Synergistic
- Antagonistic

Original	Enabled	8.26% ± 0.5%	—	—	1088 / 13167
Top high-confidence winners. Run a follow-up experiment »					
Combination 11	**Enabled**	**11.6% ± 0.6%**	**100%**	**40.6%**	**1504 / 12947**
Combination 7	**Enabled**	**10.3% ± 0.6%**	**100%**	**24.0%**	**1340 / 13073**
Combination 3	**Enabled**	**9.80% ± 0.6%**	**99.7%**	**18.7%**	**1277 / 13025**

Media	Original	**8.54% ± 0.2%**	—	—	4425 / 51794
5 / 5	Family Image	**9.66% ± 0.2%**	100%	13.1%	4996 / 51696
	Change Image	**8.87% ± 0.2%**	92.2%	3.85%	4595 / 51790

Button	Original	**7.51% ± 0.2%**	—	—	5851 / 77858
5 / 5	Learn More	**8.91% ± 0.2%**	100%	18.6%	6927 / 77729
	Join Us Now	**7.62% ± 0.2%**	73.5%	1.37%	5915 / 77644


OBAMA'08
CHANGE
WE CAN BELIEVE IN
JOIN THE MOVEMENT
Email Address
Zip Code
LEARN MORE
PAID FOR BY OBAMA FOR AMERICA
CONTINUE to WEBSITE

Section 3: A/B test or MVT

Interaction between factors

- Large interactions between factors are actually rarer than most people believe
- MVT without interaction can be thought of as running multiple A/B tests in parallel
- Ask yourself: how important is it to test interaction?

Section 3: A/B test or MVT

Bold bets and very different design

- MVT can lead to local maximum
- Try some bold bets and very different designs (A/B testing)

Section 3: A/B test or MVT

From the commentaries

- Facebook’s Protect and Care team
  Jena Cummiskey

- … letting two designers come up with very different designs and then testing them head to head … reminds me of the parallel prototyping ... I’d expect that different people could really increase the diversity of designs.
  Matt Erhart

Section 1: Overall Evaluation Criterion

Section 2: Ramp up and auto-abort

Section 3: A/B test or MVT

The Facebook Experiment

Adam D.I Kramer, Jamie E. Guillory, Jeffrey T. Hancock

Discussion

Your friend posts a picture on Facebook. He is having dinner in Paris backdropped with the Eiffel tower.

What would your response be?

What did the experiment want to prove?

Emotional Contagion

- Emotional states can be transferred to others
- Occurs outside of in-person interaction between individuals
- Nonverbal cues are not strictly necessary
- No ‘Shared Experience’ controversy

Experiment details

Modifying news feed

- Users who viewed Facebook in English
- Two parallel experiments were conducted:
  - exposure to friends’ *positive* emotional content reduced
  - exposure to friends’ *negative* emotional content reduced
- 4 groups : User group selection based on User ID
- Positive/negative posts determined by LIWC software

Findings

- Others emotions influence ours
- Non-verbal cues are not necessary
- Withdrawal effect
- Cross-emotional contagion absent
- Online messages affect offline behavior
- Effect was small

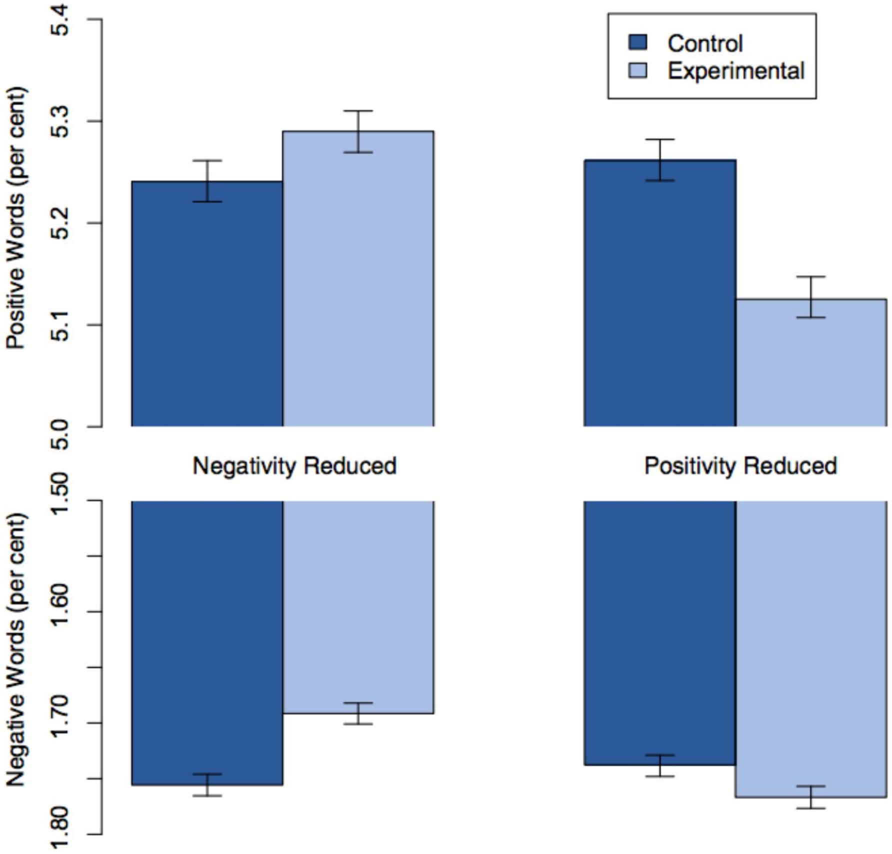


Fig. 1. Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

Why is this study important? What do we learn about web experiments?

Criticism - Unethical

Furor Erupts Over Facebook's Experiment on Users

Almost 700,000 Unwitting Subjects Had Their Feeds Altered to Gauge Effect on Emotion

- Affected user behavior
- No user consent
- The study 'harmed' participants
- Not *observational* but *experimental*

Debate

2 mins

The study is ethical, because the effect size was small

Break into groups of three.

The groups on my **left** must argue why the study **is** ethical.
The groups on my **right** argues why it is **not**.

Support

Many researchers published articles in favor of the study

In defense of Facebook

Stop complaining about the Facebook study. It's a golden age for research

The Test We Can—and Should—Run on Facebook

Ethics

Effect size **is** small

- Shifts user's own emotional word use by *two hundredths of a standard deviation*
- Facebook **removed** content; **did not add** content to **induce** behavior
- Controlled experiments are **always** being run by Facebook, Google, Twitter
  "When you use a service you don't pay for, you are not the customer, you are the product"

Problems with the experiment

- Fewer positive words produced does not mean that the user's **actual** mood was affected

- use of positive or negative words does not represent user's current emotional state

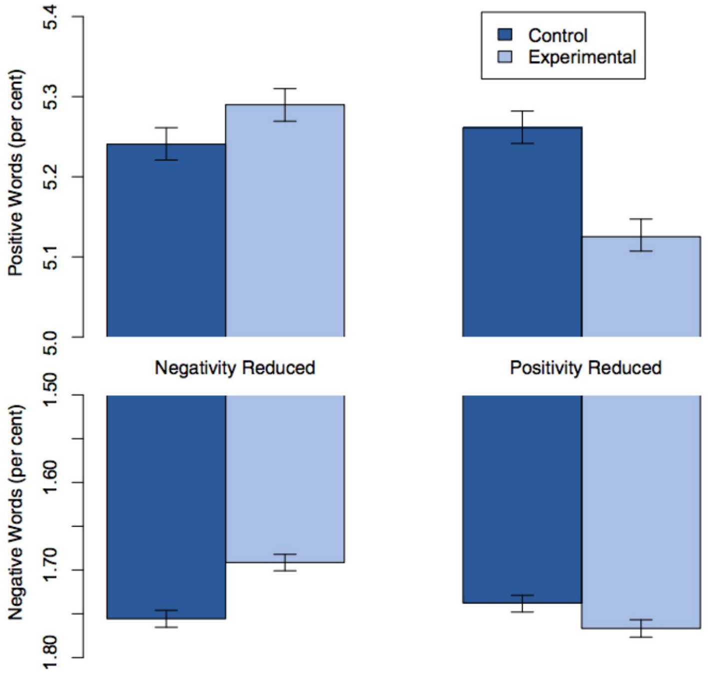


Fig. 1. Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

Problems with the experiment

Using Linguistic Inquiry and Word Count application

Consider two sentences:

> “I am not happy.”
>
> “I am not having a great day.”

LIWD score : +2 for positive (because of the words “great” and “happy”)
+2 for negative (because of the word “not” in both texts)

Actual score should be +2 on the negative scale, and 0 on the positive scale

Support for experiment by researchers

Future research will be affected

“Facebook is effectively engineering the public”

Scientific community’s access to one of the largest and richest sources of data on human behavior decreased

“amazing new platform for social science research - *companies like Facebook actually have a moral obligation to conduct such research*”

Less public visibility of experiments

Goodbye, Google

Douglas Bowman

Design at Google

Reliance on data

- Billions of shareholders at stake
- Millions of users
- Design decisions on the basis of A/B testing:
    - Reduce design decision to a simple logic problem
    - Launch if data in your favor
- No daring design decisions can be taken - testing 41 shades of blue for toolbar on Google pages

Douglas Bowman

Data, Not Design, Is King in the Age of Google

Visual Design Lead, Google - May 2006 – March 2009

First visual designer at Google

Quit Google to join Twitter as Creative Director

Greater opportunity to shape the look and feel of Twitter

“Using data is fundamental to what we do,” Mr. Bowman said. “But we take all that with a grain of salt. Anytime you make design changes, the most vocal people are the ones who dislike what you’ve done. We don’t just throw the numbers in a spreadsheet.”

Discussion

Kahavi says that **data trumps intuition** and Bowman believes in **daring design decisions**.

Are there certain situations for which **A/B testing** is always better than **hiring smart designers**, or vice versa? Why?

Commentaries

“I wonder if a designer could be trained in these kinds of factors and develop an ability to accurately predict interaction. That would be a useful skill but it’s not clear it could be explicated training.”

- Matt

“... automate creation and experimenting for system changes. I think it would be amazing if one day all we needed to do was feed an AI system a set of kinds of design changes for an interface, and that system would automatically generate controlled experiences, iterate, and learn to slowly begin changing interfaces completely on its own based on confidence thresholds.”

- Jesse

Thank you!