Online Advertising Incrementality Testing

Practical Lessons, Paid Search And Emerging Challenges

Joel Barajas, Narayan Bhamidipati, James G. Shanahan

April, 2022

Who we are ...

Joel Barajas
Sr Research Scientist
Amazon, Marketing Measurement*,
Sunnyvale, CA, USA

Linkedin

Narayan Bhamidipati
Sr. Director, Research
Yahoo! Research,
Sunnyvale, CA, USA

Linkedin

James G. Shanahan
Church and Duncan Group Inc
UC Berkeley, CA, USA

Linkedin

*Work done while employed at Yahoo! Research.

Tutorial Parts

1. The basics: context and challenges
2. Incrementality Testing: concepts, solutions and literature
3. From concept to production: platform building, challenges, case studies
4. Deployment at Scale: test cycle and case studies
5. Emerging trends: identity challenges, industry trends and solutions

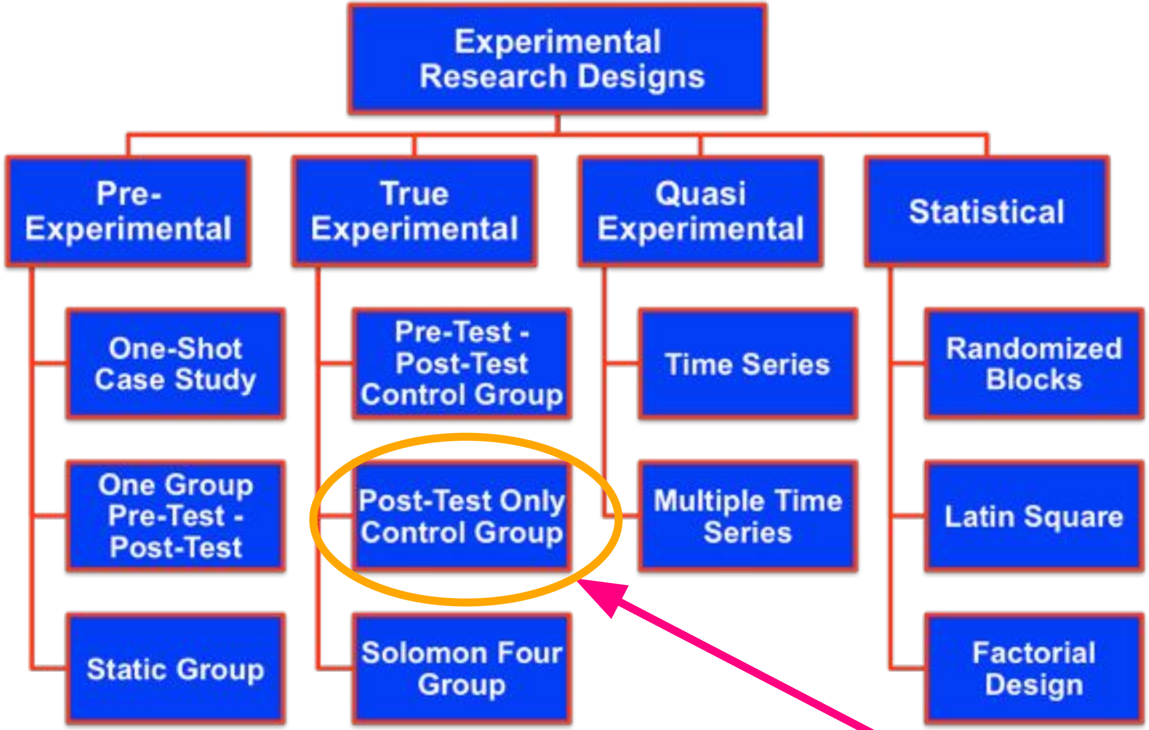
Part 2

Incrementality Testing: concepts, solutions and literature

yahoo!

ECIR 2022

Randomized Experiments: Taxonomy (link)



A/B testing
Incrementality testing

Incrementality Testing in a Nutshell

Goal:

Find Aggregate Effect of Marketing Spend

Randomized unit:

Users (our best notion)

Intervention:

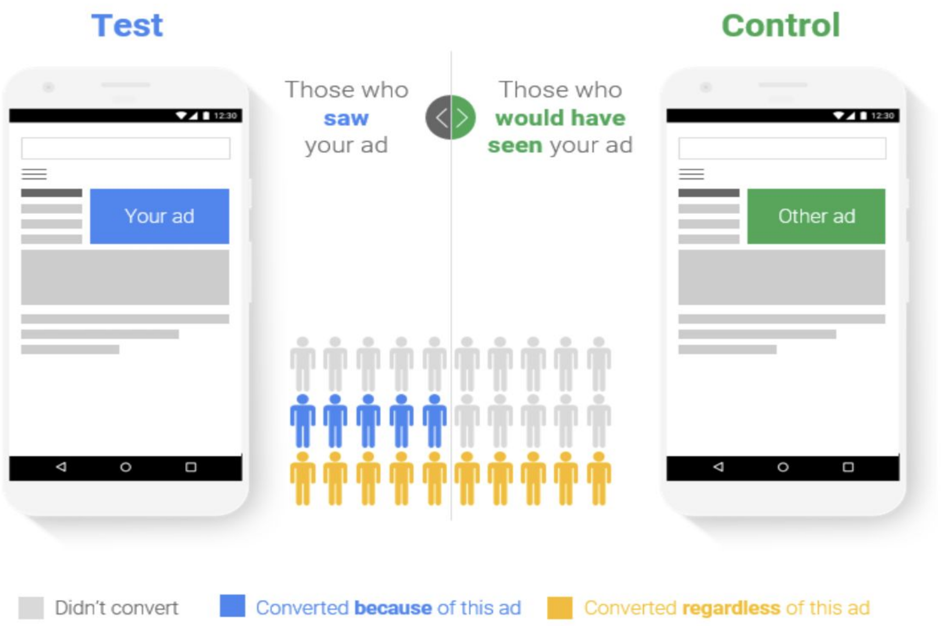
Marketing Spend leading to ad delivery

Control:

No marketing ads

Metrics:

Converter Lifts, Cost per incremental converter/conversions (CPIA), among others



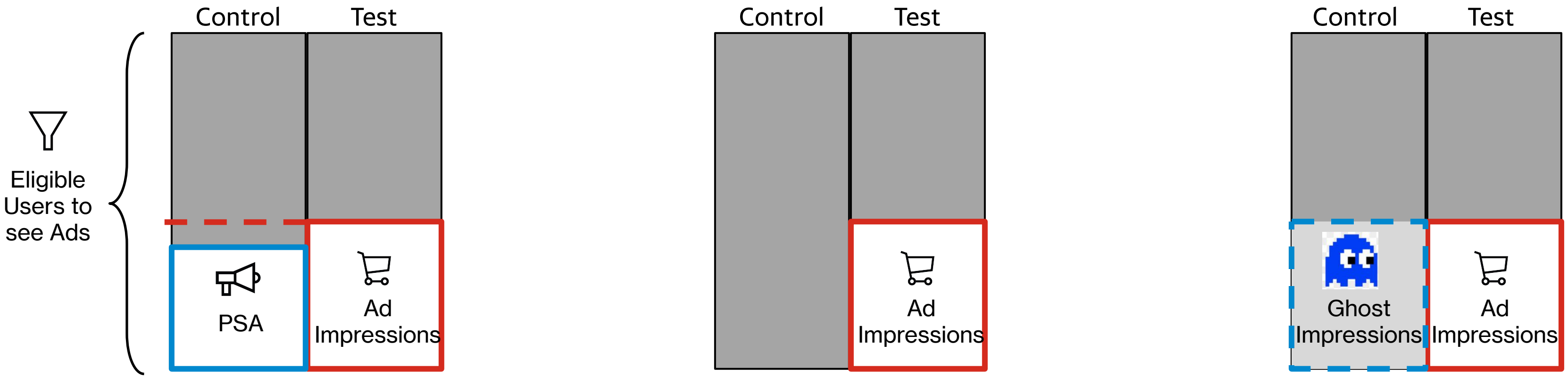
If this is *just* an A/B test, why we need more?

Short Answer: The control experience is no Ads thus it is difficult to identify control users

Experimentation Typical Designs for incrementality testing

Lewis et al. (2011), Barajas et al. (2016), Johnson et al. (2017), Barajas and Bhamidipati (2021)

Key Challenge: Identify would-be (counterfactual) impressed users



Case 1: Placebo PSA

Case 2: Intent to Treat

Case 3: Ghost Ad Approach

Potential misalignment in user groups

Diluted effect design reducing the test power

Impressed and "ghost" impressed users are compared

yahoo!

Experimentation Typical Designs for incrementality testing

Lewis et al. (2011), Barajas et al. (2016), Johnson et al. (2017), Barajas and Bhamidipati (2021)

Randomized Design	Selection Bias Risk	Design Precision Level	Engineering Cost
PSA Testing	**High** - targeting knows user groups	**ad impression**	low
Intention to Treat	**Minimal** - blind to targeting	ad opportunity	medium
Ghost Ads	**Minimal** - blind to targeting	**ad impression**	**high**

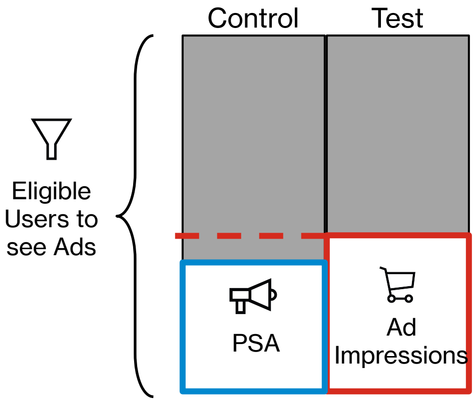
Ghost Ads provides the **highest precision without selection bias** but with the **highest Engineering Cost**

yahoo!

Placebo Public Service Announcements (PSA) Based Testing

Lewis et al. (2011)

Run Placebo Campaigns to Replicate User Targeting

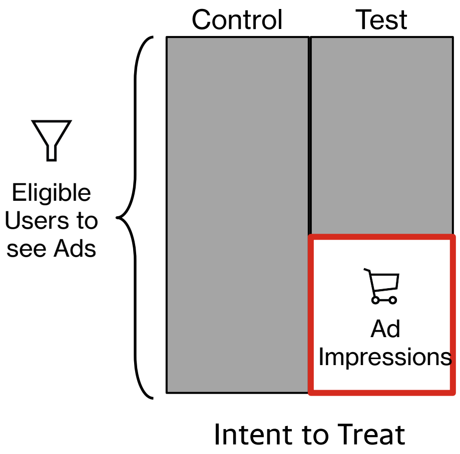


Placebo PSA

Potential misalignment in user groups

- It requires setting up **two targeting models** paying the costs of PSA ads
  - The targeting model in control **will NOT get the same feedback**
  - Introducing **selection bias** after a few weeks of testing
- Fundamental issue: **it is not double blind design**
  - It is **not blind to the targeting engine** as treatment administrator

Intent to Treat Based Testing
Barajas et al. (2016)



Diluted effect design reducing the test power

Set Aside a Group of Users without Ads

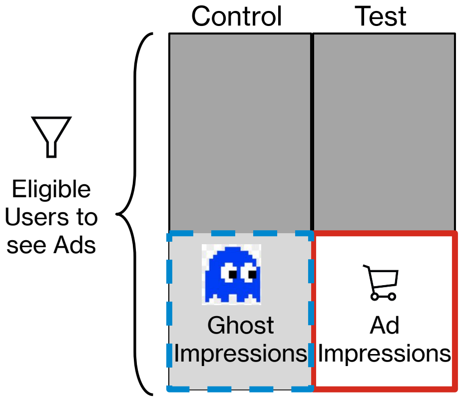
- It requires a **qualifying event that is NOT influenced by the treatment** to filter the users in the analysis
  - **Visiting users** to publishers' pages
  - Filter users based on **target segments**

- Fundamental issue: it dilutes the effect greatly **decreasing the statistical power**
  - Since **we can not identify the users who would have seen the ad** in the control group, all users need to included in the estimation

Ghost Ads Based Testing

Johnson et al. (2017), Barajas and Bhamidipati (2021)

Identify and Log counterfactual "ghost" impressions



Ghost Ad Approach

Impressed and "ghost" impressed users are compared

- It requires **engineering effort to hold out control users** and log their ghost impressions
  - The hold out point is equivalent to **exposure logging events** in A/B experimentation platforms
- It provides the **same statistical power** as PSA based testing but truly **double blind experiment** design
  - **Ad networks:** it provides **post-auction** user randomization
  - **Third-party exchanges:** this precision is achieved by **matching bid prices** between the hold out ad and the ad sent to the exchange

yahoo!

ECIR 2022

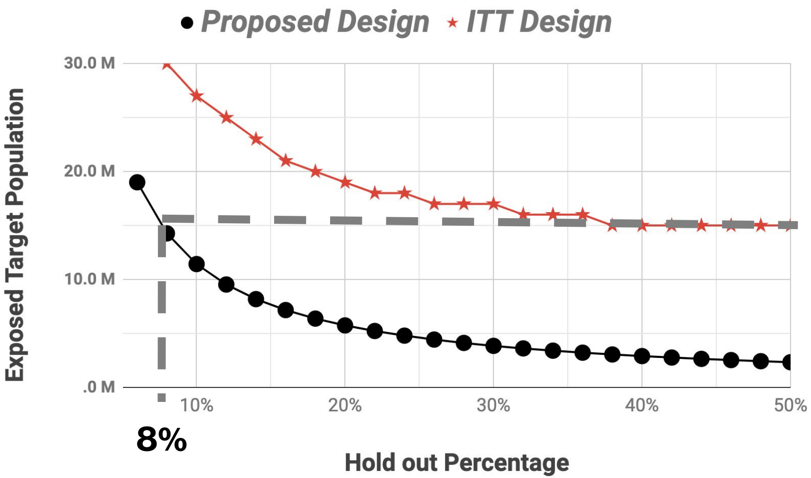
Results: Increased Precision Benefits

Minimum Detectable Lift	Converter Rate (Control Group)		Converter Rate (Treat Group)		Exposed Users Needed		**Design Gain**
	Ad Targeted	No Ad Targeted	Ad Targeted	No Ad Targeted	Design Users	ITT	
15%	0.135%	0.05%	0.155%	0.05%	1.28M	8M	**84%**
10%	0.135%	0.05%	0.149%	0.05%	2.80M	12M	**76%**
5%	0.135%	0.05%	0.142%	0.05%	11.40M	27M	**58%**
3%	0.135%	0.05%	0.139%	0.05%	31.67M	49M	**35%**

Precision gain between proposed design and literature ITT design with ghost ads for ad exchanges.
Control group size: 10%. Reachable bidded users: 100M. Confidence level: 95%.

Ghost Impressions design leads to up to 84% less users needed to achieve a statistical significant read.

ITT (Case 2) at 50% control size power is reached at 8% of our design.



Control converter rate: 0.135%. Minimum detectable lift: 5%. Reachable users: 100M. Confidence level: 95%.

Causal Inference Estimation

Review of Causal Inference Framework

Causal Inference Frameworks: Potential Outcomes
Rubin (2005)

Everything is written in terms of experiment units, treatments and potential Outcomes

- The causal inference problem is defined by **hypothesizing a counterfactual universe** without the treatment and comparing the user responses in both universes
  - This framework **separates the causal setup from the inference** problem

- The Statistical Inference problem is defined as a **missing value problem**
  - It provides a **fundamental framework to integrate experiment blocking** and to account for biases in the data collection

yahoo!

ECIR 2022

Causal Inference Frameworks: Potential Outcomes

Frangakis and Rubin (2002), Imbens and Rubin (1997)

Finding Average Treatment Effects requires careful handling of conditional user features

- The **average treatment effect** is the target statistic to attribute a causal difference
  - By definition, **the average response over the treatment units**, eg users.

- User features fall into: **pre-treatment and post-treatment** feature groups
  - Filtering users, eg finding *conditional treatment effects*, requires **testing the variables for post-treatment bias**

Causal Inference Frameworks: Potential Outcomes

Frangakis and Rubin (2002), Imbens and Rubin (1997)

In A/B testing, user treatment assignment is *ignorable*

- Since **treatment assignment is *random*, they are ignorable** allowing for a straight mean difference statistical test
  - The **average must be taken over users**, eg conversions per user, NOT impressions, NOT visits, or any other events
  - If stratified sampling is deployed, ie experiment blocking, the **stratifying features must be included in the inference** since they are NOT ignorable
  - When effects on multiple metrics are analyzed they must be **estimated in isolation** without conditioning users on these metrics (post-treatment variables)

Causal Estimation and Metrics

Average Treatment Effect (ATE) and Lift:

$Z_i = 0$ for the control and $Z_i = 1$ for the test group.

$$ATE = E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0] \qquad CR\ lift = ATE / E[Y_i | Z_i = 0]$$

Leveraging Central Limit Theorem:

$$ATE \sim N(\bar{Y}_1 - \bar{Y}_0, \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0})$$

$N(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2

Metric: Y = Converters

$$n_Z = \#\ of\ users\ in\ group\ Z$$

$$\bar{Y}_Z = \frac{\#\ converters\ in\ group\ Z}{\#\ of\ exposed\ users\ in\ group\ Z}$$

$$S_Z^2 = \bar{Y}_Z * (1 - \bar{Y}_Z)$$

Metric: Y = Conversions

$$\bar{Y}_Z = \frac{\#\ conversions\ in\ group\ Z}{\#\ of\ exposed\ users\ in\ group\ Z}$$

$$\bar{Y^2}_Z = \frac{(\#\ conversions\ in\ group\ Z)^2}{\#\ of\ exposed\ users\ in\ group\ Z}$$

$$S_Z^2 = \bar{Y^2}_Z - (\bar{Y}_Z)^2$$

Cost per Incremental Converter

$$CPiA = \frac{marketing\ spend\ (\$)}{ATE \times (\#\ of\ users\ in\ test\ group)}$$

Incremental Return on Ad Spend

$$iROAS = \frac{ATE \times (\#\ of\ users\ in\ test\ group)}{marketing\ spend\ (\$)}$$

!

Potential Outcomes Causal Model:
Randomized Units must be aligned

Ignorable Treatment Assignment to Features:
No stratification or blocking necessary in the estimation

“In Theory There Is No Difference Between Theory and Practice, While In Practice There Is”

We’ll review execution in the next part of the tutorial....