Field Experiments on Online Advertising

David Reiley
(based on research with Randall Lewis)
Tech Cities conference
March 2015
The authors would like to acknowledge Yahoo! for the support we received in doing this research there.

All material contained here has been approved by Yahoo! for public presentation.
Advertising’s effects on sales have always been very difficult to measure.

“Half the money I spend on advertising is wasted; the trouble is I don't know which half.”

-John Wanamaker
(Department store merchant, 1838-1922)
That was 100 years ago.

• John Wanamaker died in 1922.
  – Before penicillin, quantum mechanics, radar, lasers, jet engines, space travel, the Internet…

• How is it that we have made such dramatic progress in the natural sciences, yet still don’t know much about advertising effectiveness?
Come on, are things really so bad?

• We have tons more data now than we did 100 years ago.
  – Nielsen, comScore, surveys, purchase data, online clicks and conversions.
• Statistical modeling has become more sophisticated.
• But fundamentally, observational data only reveals correlation, not causation.
Why is ad effectiveness hard to measure?

- Revenue
- Advertising
- Spend
Advertise more at Christmas. (★)

- Revenue
- Advertising
- Spend
Plenty of reasons why Media Mix Modeling can give us wrong answers.

- MMM uses statistical regressions of aggregate sales versus aggregate advertising.
  - No matter how complex the model, the fundamental idea is to identify variation, as in the preceding plots.
- Key question: what causes advertising to vary over time?
  - Correlation is not the same as causation.
  - Reverse causality possible.
  - Some third variable may cause both ads and sales.
A recent Harvard Business Review article illustrates the state of the art in measuring online ad effectiveness.

- Technique: Compare those who saw an online ad with those who didn’t.
What’s wrong with comparing exposed versus unexposed users?

- Potential problem: the two samples do not come from the same population.
- Example: Who sees an ad for eTrade on Google?
  - Those who search for “online brokerage” and similar keywords.
  - Does the ad actually cause the difference in sales?
- Correlation does not mean causality.
We have just seen two ways for observational data to provide inaccurate results.

- Aggregate time-series data
  - Advertising doesn’t vary systematically over time.

- Individual cross-sectional data
  - The types of people who see ads aren’t the same population as those who don’t see ads.
  - Even in the absence of any ads, they might well have different shopping behavior.
  - Becomes even more true as advertising becomes more highly targeted.
An experiment is the best way to establish a causal relationship.

- Systematically vary the amount of advertising: show ads to some consumers but not others.
- Measure the difference in sales between the two groups of consumers.
- Like a clinical trial for a new pharmaceutical.
- Almost never done in advertising, either in online or traditional media.
  - Exceptions:
    - Direct mail.
    - Split-cable TV experiments (IRI BehaviorScan)
Our understanding of advertising today resembles our understanding of physics in the 1500s.

- Do heavy bodies fall at faster rates than light ones?
- Galileo’s key insight: use the experimental method.
- Huge advance over mere introspection or observation.
At Yahoo! Labs, we put the experimental method to good use.

• Several different clients have agreed to careful experiments using a control group:
  – Three retailers (Offline sales data!)
  – Several online service providers
  – Internal Yahoo! properties

• A variety of outcomes can be measured:
  – Online sales or other conversions
  – Offline sales, in special cases
  – Survey questions (brand affinity, etc.)
  – Online searches
Lewis and Reiley (2014): American retailer measures economically large effects of online ads.

- This retailer keeps careful records, attributing >90% of in-store purchases to the correct individual customer.
- We found 1.6 million customers who matched (name and address, either email or snail-mail) between the databases of the retailer and Yahoo!
- 80% of matched customers assigned to the treatment group
  - Targeted with retail-image ad campaigns from the retailer
- 20% assigned to the control group
  - Do not see these retailer ads.
- Ad campaigns are “Run of Network” on Yahoo!
- Following the online ad campaigns, we received both online and in-store sales data: for each week, for each person
  - Third party de-identifies observations to protect customer identities
  - Retailer disguises all sales amounts (R$) with a scalar multiple of USD
Descriptive statistics for Campaign #1 indicate valid treatment-control randomization.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Female</td>
<td>59.5%</td>
<td>59.7%</td>
</tr>
<tr>
<td>% Retailer Ad Views &gt; 0</td>
<td>0.0%</td>
<td>63.7%</td>
</tr>
<tr>
<td>% Yahoo Page Views &gt; 0</td>
<td>76.4%</td>
<td>76.4%</td>
</tr>
<tr>
<td>Mean Y! Page Views per Person</td>
<td>358</td>
<td>363</td>
</tr>
<tr>
<td>Mean Ad Views per Person</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>% Ad Impressions Clicked (CTR)</td>
<td>-</td>
<td>0.28%</td>
</tr>
<tr>
<td>% Ad Viewers Clicking at Least Once</td>
<td>-</td>
<td>7.21%</td>
</tr>
</tbody>
</table>
Experimental differences show a positive increase in sales due to ads.

<table>
<thead>
<tr>
<th></th>
<th>During Campaign</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(2 weeks)</td>
<td>(2 weeks)</td>
</tr>
<tr>
<td></td>
<td>Mean Sales/Person</td>
<td>Difference</td>
</tr>
<tr>
<td>Control:</td>
<td>R$ 1.84</td>
<td>-R$ 0.10</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Treatment:</td>
<td>1.89</td>
<td>-R$ 0.04</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
</tbody>
</table>

- 95% C.I. for treatment is R$0.05 ± 0.07.
- Treatment effect on the treated: R$0.085 ± 0.11.
Suppose we had no experiment, and just compared spending by those who did or did not see ads.

<table>
<thead>
<tr>
<th></th>
<th>During Campaign</th>
<th>(2 weeks)</th>
<th>Mean Sales/Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control:</td>
<td>R$ 1.84</td>
<td>(0.03)</td>
<td></td>
</tr>
<tr>
<td>Treatment:</td>
<td>1.89</td>
<td>(0.02)</td>
<td></td>
</tr>
<tr>
<td><strong>Exposed to Retailer’s Ads:</strong></td>
<td><strong>1.81</strong></td>
<td>(0.02)</td>
<td></td>
</tr>
<tr>
<td>[64% of Treatment Group]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Not Exposed to Retailer’s Ads:</strong></td>
<td><strong>2.04</strong></td>
<td>(0.03)</td>
<td></td>
</tr>
<tr>
<td>[36% of Treatment Group]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- We would conclude that ads decrease sales by R$0.23!
- Not comparing apples to apples here.
Pre-campaign data emphasize that the non-experimental sales differences have no causal relationship to the ad exposures.

<table>
<thead>
<tr>
<th></th>
<th>Before Campaign (2 weeks)</th>
<th>During Campaign (2 weeks)</th>
<th>Difference (During – Before)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Sales/Person</td>
<td>R$ 1.95</td>
<td>R$ 1.84</td>
<td>-R$ 0.10</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.03)</td>
<td>(0.05)</td>
</tr>
<tr>
<td><strong>Treatment:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Sales/Person</td>
<td>1.93</td>
<td>1.89</td>
<td>-R$ 0.04</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Exposed to Retailer’s Ads:</td>
<td>1.81</td>
<td>1.81</td>
<td>0.00</td>
</tr>
<tr>
<td>[64% of Treatment Group]</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td></td>
</tr>
<tr>
<td>Not Exposed to Retailer’s Ads:</td>
<td>2.15</td>
<td>2.04</td>
<td>-R$ 0.10</td>
</tr>
<tr>
<td>[36% of Treatment Group]</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
</tbody>
</table>

- People who browse enough to see ads also have a lower baseline propensity to purchase from the retailer.
- Potential mistake solved with experiment.
Ad exposures appear to have prevented a normal decline in sales during this time period.

<table>
<thead>
<tr>
<th></th>
<th>Before Campaign (2 weeks)</th>
<th>During Campaign (2 weeks)</th>
<th>Difference (During – Before)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Sales/Person</td>
<td>Mean Sales/Person</td>
<td>Mean Sales/Person</td>
</tr>
<tr>
<td><strong>Control:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R$ 1.95 (0.04)</td>
<td>R$ 1.84 (0.03)</td>
<td>-R$ 0.10 (0.05)</td>
</tr>
<tr>
<td><strong>Treatment:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposed to Retailer’s Ads:</td>
<td>1.81 (0.02)</td>
<td>1.81 (0.02)</td>
<td>R$ 0.00 (0.03)</td>
</tr>
<tr>
<td>[64% of Treatment Group]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Exposed to Retailer’s Ads:</td>
<td>2.15 (0.03)</td>
<td>2.04 (0.03)</td>
<td>-R$ 0.10 (0.04)</td>
</tr>
<tr>
<td>[36% of Treatment Group]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Control-group sales fall.
- Unexposed treatment-group sales fall.
- Treated-group sales stay constant.
Our difference-in-difference estimate yields a statistically and economically significant treatment effect.

- Estimated effect per customer of viewing ads:
  - Mean = R$ .102, SE = R$ .043
- Estimated sales impact for the retailer:
  - R$83,000 ± 70,000
    - 95% confidence interval.
    - Based on 814,052 treated individuals.
    - Compare with cost of about R$25,000.
    - 325% increase in revenue relative to cost.
- Note the wide confidence interval. But it’s actually much narrower than that for a “successful” IRI BehaviorScan test.
  - more like R$83,000±190,000.
What happens after the two-week campaign is over?

• Positive effects during the campaign could be followed by:
  – Negative effects (intertemporal substitution)
  – Equal sales (short-lived effect of advertising)
  – Higher sales (persistence beyond the campaign)

• We can distinguish between these hypotheses by looking at the week following the two weeks of the campaign.
We now take a look at sales in the week after the campaign ends.

• Previous two-week estimate:
  – R$0.102 (0.043) per person.
• Estimate for third week:
  – R$0.061 (0.024) per person.
  – As large as the effect per week during the campaign.
• Including the third week, the total impact of the ads becomes:
  – R$135,000 ± 85,000
  – Compared with cost of R$25,000.
• Extending out five weeks, the total looks as high as R$250,000 ± 190,000 (compared with cost of R$33,000).
Next we estimate separate effects for the effect on offline and online sales.

<table>
<thead>
<tr>
<th>Viewed Ads</th>
<th>Total Sales</th>
<th>Offline Sales</th>
<th>Online Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewed Ads</td>
<td>R$ 0.166</td>
<td>R$ 0.155</td>
<td>R$ 0.011</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

- Three-week sales lift per person.
- 93% of the total effect occurs in stores!
Do we capture the effects of ads by measuring only clicks? No.

<table>
<thead>
<tr>
<th>Viewed Ads</th>
<th>Total Sales</th>
<th>Offline Sales</th>
<th>Online Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R$ 0.166</td>
<td>R$ 0.155</td>
<td>R$ 0.011</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Did Not Click [93%]</td>
<td>R$ 0.139</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clicked [7%]</td>
<td>R$ 0.508</td>
<td>R$ 0.215</td>
<td>R$ 0.292</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.16)</td>
<td>(0.04)</td>
</tr>
</tbody>
</table>

- 78% of the lift comes from viewers who never clicked!
- With an experiment, no attribution model is required.
The effect of online display ads varies with browsing behavior.

The largest effect of the advertising was on customers who browsed enough to see between 1 and 100 ads.
The increase appears to consist of about $\frac{1}{4}$ increase in transactions, $\frac{3}{4}$ increase in basket size.

- **Prob(transaction) increases by 0.1% (0.05%)**
  - Baseline amount = 6.5%
  - Percentage increase = 1.5%.

- **Mean basket size increases by R$1.75 (0.74)**
  - Baseline amount = R$41.
  - Percentage increase = 4.2%

- Both effects are statistically significant at the 5% level.
Wasted half? Older customers’ purchases responded more than younger ones.

- For this campaign, 38% of the total effect comes from the 6% of consumers over 65 years of age.
We designed a second study to measure the impact of frequency.

• See Johnson, Lewis, and Reiley (2015)
• This time, we have 3 million matched users.
• Two campaigns in two weeks.
• Three equal-sized treatment groups:
  – Control (no ads)
  – Half frequency (17 impressions/person on average)
  – Full frequency (34 impressions/person on average)
• This time, we deliver Y! house ads as control impressions.
  – Mark hypothetical views of the Control group.
  – Also mark views the Half group would have seen in the Full group.
• Again we see in-store and online transaction data for each customer during the experiment.
• Real US dollars this time.
Again, randomization looks good.

The Age Distribution is Identical for All Three Groups

![Graph showing age distribution for three groups: 'None', 'Half', 'Full'. The distribution is centered around the 30-40 age group with a peak at around 35 years old. The frequency counts show a consistent pattern across all groups.]
Frequency has surprisingly high marginal impact when going from 17 to 34 ads per person.

<table>
<thead>
<tr>
<th>Purchases During 2 Weeks</th>
<th>Control</th>
<th>Half</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean purchase amount</td>
<td>$17.62</td>
<td>$17.98</td>
<td>$18.22</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.17)</td>
<td>(0.17)</td>
</tr>
<tr>
<td>Difference from control</td>
<td>0.36</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.24)</td>
<td></td>
</tr>
</tbody>
</table>

- Doubling the frequency increases sales by 50% more.
- Diminishing returns not as high as we might have expected!

*We include 60% of users, all those who viewed either a treatment or a control ad.
*This aggregates effects of both one-week campaigns. Second week showed much larger effects than the first.
The impact of the ads is greatest for those who live within two miles of a retail store.

- $0.50 avg effect for all customers: 7X ROAS.
- $3.00 avg effects for those within 2 miles of a store: 36X ROAS!
At Yahoo!, we can also measure the increase in searches due to display ads.

- In the treatment group, 1300 brand searches take place within 10 minutes of exposure to an impression in this ad campaign.

* Histogram bins are 5 seconds wide.
Without an experiment, we would have overestimated the effects.

- True incremental impact: 560 searches, not 1314.
Correlated online behaviors can lead to overestimates of the effects of advertising.

- Activity bias: People who are doing an activity online are much more likely to be doing other online activities at around the same time.
- Lewis, Rao, and Reiley (2011), “Here, There, and Everywhere,” documents several examples of activity bias:
  - Impact of display ads on search queries for an advertiser’s brand.
  - Impact of display ads on online conversions (new account applications).
  - Impact of Y! video ads shown on Amazon Mechanical Turk for Y! page views.
- In each case, failing to use an experiment gives us an overestimate of true causal effects.
Conclusion: With data, quality is more important than quantity.

- No amount of Big Data can fix the problem that correlation is not the same as causation in observational data.
  - Spurious correlation of time trends
  - Omitted variables (Xmas)
  - Reverse causality
  - Sample selection (eTrade ads on Google)
- Field experiments create variation in X that is systematically uncorrelated with other types of variation in Y. That gives us causal inference.
- See my website for:
  - My research manuscripts
  - Link to my course on field experiments at Berkeley