Advertising Incrementality Measurement using Controlled Geo-Experiments: The Universal App Campaign Case Study*

Joel Barajas†
Yahoo! Research, Verizon Media
Sunnyvale, CA
joel.barajas@verizonmedia.com

Tom Zidar‡
Data Science, Uber
San Francisco, CA
tzidar@uber.com

Mert Bay
Marketing Data Science, Uber
San Francisco, CA
mert@uber.com

ABSTRACT

Measuring the incrementality (effectiveness) of advertising is a critical task for advertisers for financial planning and optimal budget allocation among different online channels. Recent literature has consistently recommended the use of experiments to estimate advertising effectiveness reliably. Closed Ad networks often prevent advertisers from having direct access to user-level traffic for business privacy reasons. As a result, running experiments by randomizing users is rarely feasible for advertisers. We present a controlled experiment design and an effect estimation framework focused on advertisers' side by leveraging geo-targeted spend interventions at market-level. Our method is based on the selection of the best pair of markets for testing, conditional on a pre-determined effect estimation method, preventing any model tuning bias. We use a Bayesian structural time series to predict the treatment conversions counterfactual based on the observed control market conversions. We present the results of a field experiment of a Universal App Campaign (UAC), a recent mobile ad campaign format. We find evidence that this advertising format causes incremental conversions, despite the limited campaign customization options. We measure a 6.57% decrease of conversions (statistically significant) when UAC spend is suspended. To our knowledge, our work is one of the earliest studies that successfully measures the incremental value of UAC with controlled experiments.

ACM Reference Format:
Joel Barajas, Tom Zidar, and Mert Bay. 2020. Advertising Incrementality Measurement using Controlled Geo-Experiments: The Universal App Campaign Case Study. In *Proceedings of (ADKDD 2020).* ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Advertisers often run a portfolio of media channels, including programmatic native advertising, social advertising, sponsored search, video advertising, among others. In these heterogeneous ecosystems, relying on last-touch attribution poses significant issues because of the bias towards channels closer to conversion in the funnel (demand capture channels), e.g. sponsored search, versus demand generation channels, e.g. display advertising. Thus, measuring the incrementality of advertising spend is critical for advertisers' financial planning and optimal budget allocation [14].

Previous literature in the evaluation of online campaigns has focused mainly on the effectiveness of display advertising [11], how ad targeting incentives play a role in the evaluation [2], and the difficulty of measuring the effect of this advertising format [15]. Other studies have evaluated paid search [12], and social advertising [9]. Most of these results rely on the ability to randomize users and detailed user tracking data in all treatment groups. Another stream of research to solve this problem is running user-level observational-based methods, including propensity score-based methods [16], from logged data post-campaign. However, Gordon *et al.* concluded that these methods, even in the presence of rich user-level data, do not provide a reliable estimate of the advertising effectiveness [9]. The most widely accepted approaches by the research community, as well as by the ad tech industry, to measure advertising incrementality are to use controlled experiments [4, 8, 11].

Evaluating campaigns for optimal budget allocation implies measuring the incremental value using advertiser side data and interventions. Product conversions (user responses) are generally well-identified within markets by advertisers. However, in the online advertising ecosystem, ad networks typically manage user-level targeting and ad delivery. Thus, advertisers are often prevented from having direct access to user level traffic, especially when campaign ads are not displayed. As a result of this limited access, running incrementality experiments by randomizing users is rarely feasible outside the ad network. Therefore, advertisers are left with aggregated time series response signals and spend intervention levers, which can be targeted at the market level. This time-series data lead to the deployment of Media Mix Models, which suffer from the lack of rigorously controlled experiment results in the estimation of the advertising budget effectiveness [7].

Designed Market Area (DMA) based targeting in the US is a testing strategy that uses aggregate time series and counterfactual prediction [4, 13]. DMAs are geographical regions in the US designed based on marketing similarities[1]. Kerman *et al.* and Blake *et al.* recommend to randomize DMAs, or any other geo segmentation outside the US, aggregate the observed responses overall regions, and perform a pre/post-intervention analysis based on difference-in-differences [4] or Bayesian Structural time series [13].

*This work represents the views of authors only.
†Work done while the author was employed at Uber
‡Work done while the author was employed at Uber

[1]Nielsen DMA Regions. https://www.nielsen.com/intl-campaigns/us/dma-maps.html

ADKDD 2020, 2020

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.1145/1122445.1122456

One fundamental limitation of this design is the need to pause the advertising strategy in a given country for the test. Since the audience for an advertising channel format (e.g. social advertising) is likely to overlap with other media channels in the media mix (e.g. programmatic display), the risks of channel cannibalization might bias the experiment. Pausing the advertising strategy in the entire country is rarely feasible in ongoing advertising spend planning.

Blake *et al.* address geographical spillovers among DMAs, and the difficulties from merely randomizing DMAs leading to a simple correlation analysis to identify the best control group, i.e., not at random. Also, the experiment required close to two-thirds of DMAs. However, the authors do not address the bias of other media channels spend, and the spend levels heterogeneity of the markets before the experiment intervention.

1.1 Our Contribution

We introduce a controlled geo-experiment design and an effect estimation framework focussed on advertisers' needs. We model the treatment effect using a Bayesian structural framework with time series and a regression component (matched control market conversions), to predict the treatment conversions [5] within the family of synthetic control based methods [1].

We focus primarily on the controlled geo-experiment design. Thus, given the causal effect estimation method, we perform several placebo tests (i.e., A/A tests) using the experiment response signal and the typical markets for the advertiser. Based on these tests, we select the best market matched pairs, as opposed to a simple correlation analysis proposed by Blake *et al.* [4]. This process guarantees unbiased model selections or tuning post-experiment since we design the controlled experiment specifically for the user response of interest and the given causal effect estimation.

We present the results of a controlled experiment of a Universal App Campaign (UAC), a recent mobile ad campaign format[2]. In this format, the ad network manages and optimizes UAC campaigns within its properties. However, advertisers often expressed concern about limited settings and transparency. The current experiment was performed by Uber as an advertiser, where Uber cities were used as geo-market units. We provide evidence that this advertising format is effective in driving incremental customer acquisition conversions, i.e., signups. To our knowledge, our work is one of the earliest studies to be published that successfully measures the incremental value of UAC with controlled experiments.

Running market matched controlled experiments comes with a significant amount of experiment design tuning and recommendations to improve the precision of the measurement. Challenges include confounding factors such as spend in other media channels, different effects of holidays on target markets, decreased marginal effects of additional spend (diminishing returns [7]), among others. We provide detailed recommendations and discussions about the practical challenges.

2 METHODOLOGY

We propose a controlled market match based experiment design where we identify the best market pair matching. This design is based on the causal estimation of placebo tests (A/A tests) and their credible intervals. Thus, we first describe the causal estimation methodology followed by the controlled experiment design.

2.1 Causal Estimation

Given a pair of markets, we use the conversion time series in both. We consider these markets as treatment and control cells and use the control cell as a predictive variable of the treatment cell.

We model the treatment effect using a Bayesian structural framework with time series and a regression component (matched control market conversions), to predict the treatment conversions (synthetic control). We consider the following structural equation:

$$\begin{aligned} y_t^{(treat)} &= F\theta_t + x_t^{(control)T}\beta + \epsilon_t, \quad & \epsilon_t &\sim N(0,\sigma^2), \\ \theta_t &= G\theta_{t-1} + \omega_t, & \omega_t &\sim N(0,W), \end{aligned} \tag{1}$$

where t is the week index, $y_t^{(treat)}$ is the weekly number of conversions in the treatment market, $x_t^{(control)}$ is the weekly number of conversions in the control market, θ_t is the time series latent state of $y_t^{(treat)}$, F and G are the state design matrices, β is the regression parameter vector of control conversions, and σ^2, W are the variance parameters. This state-space time series modeling is flexible to integrate various seasonal and trend components based on model superposition in the definition of F and G. Different components have been modeled before in the effect estimation of online display campaigns [3]. We note the flexibility of the estimation framework and the family of model selections it covers.

We model a local linear trend, a spike-and-slab prior distribution for β, diagonal covariance matrix W, and standard non-informative conjugate inverse gamma priors of the variance parameters. The model is fitted using a Markov Chain Monte Carlo (MCMC) sampling approach for β [10], and a faster alternative to the standard forward-filtering, backward-sampling approach in state-space models [5]. The spike-and-slab prior distribution for β provides an automatic variable selection for cases when the control time series is approximately uncorrelated to the treatment series.

We consider historical conversion time series for both treatment and control groups to fit the model. Then, at the time of intervention, we predict the treatment conversions by the evolution of the time series components and the control conversion observations. Let $\Theta = \{\sigma^2, W, \beta\}$ be the model parameters to be fitted, given pre-intervention historical conversions, $D_{1:T-1} = \{y_{1:T-1}^{(treat)}, x_{1:T-1}^{(control)}\}$, where T is the time of intervention. After fitting the model, we have the posterior distribution samples Θ^s given $D_{1:T-1}$ where $s = 1, \ldots, N_s$ and N_s is the number of MCMC samples. We predict the treatment conversions after intervention as follows:

$$\hat{y^s}_{t_i}^{(treat)} = F\theta_{t_i}^s + x_{t_i}^{(control)T}\beta^s|\{D_{1:T-1}, x_{t_i}^{(control)}\} \tag{2}$$

for $\forall s \in \{s = 1, \ldots, N_s\}$ and $\forall t_i \in \{T, \ldots, T_i\}$, where t_i is the time index after intervention and T_i is the latest observation. Based on these $y^{1:\hat{N}_s}{}_{T:T_i}^{(treat)}$ posterior samples, and the actual observations for the treatment group after intervention $y_{T:T_i}^{(treat)}$, we estimate

[2]About Universal App Campaigns. https://support.google.com/google-ads/answer/6247380?hl=en

Algorithm 1 Control/Treatment Market Pair Selection

```
1: Ω: Set of Markets to consider
2: Φ: Set of placebo intervention times
3: Δ_t: Time length of historical data
4: Δ_{t_i}: Time after placebo intervention
5: for all treatment market: m ∈ Ω do
6:     for all control market: n ∈ {Ω − m} do
7:         for all intervention time: d ∈ Φ do
8:             Fit the synthetic control model of Eq 1:
9:             Find Θ^s, s = 1, ..., N_s, given {y^(m)_{d−Δ_t:d−1}, x^(n)_{d−Δ_t:d−1}}
10:            Predict ŷ^s(m)_{t_i}, ∀s ∈ {s = 1, ..., N_s} after intervention,
               ∀t_i ∈ {d, ..., d + Δ_{t_i}}
11:            Estimate Credible Intervals (CI) lift_{cum(d+Δ_{t_i})}, Eq 3
12:        end for
13:    end for
14:    n*, d* ← tightest CI that include lift_{cum(d+Δ_{t_i})} = 0
15:    Append best control/treatment/time V = {V, (m, n*, d*)}
16: end for
17: return V
```

intervention attribution change as:

$$
\begin{aligned}
y^s_{Attr(t_i)} &= y^{(treat)}_{t_i} - \hat{y^s}^{(treat)}_{t_i},\\
y^s_{cumAttr(t_i)} &= \sum_{t \in T,\dots,t_i} y^s_{Attr(t)},\\
lift^s_{cum(t_i)} &= \frac{y^s_{cumAttr(t_i)}}{\sum_{t \in T,\dots,t_i} \hat{y^s}^{(treat)}_{t_i}},
\end{aligned}
\tag{3}
$$

where $y^s_{Attr(t_i)}$ is the conversion change attributed to the treatment intervention at time t_i, $y^s_{cumAttr(t_i)}$ is the cumulative attributed conversions up to time t_i, and $lift^s_{cum(t_i)}$ is the cumulative lift up to t_i, all for the s sample. Posterior credible intervals are estimated based on the N_s effect estimate samples. For the effects of the current experiment analysis, we use the *CausalImpact* implementation in the interest of result replicability [6].

Compared to the time-based regression approach for analyzing geo experiments proposed by Kerman *et al.* in [13], which uses only the control markets in the treatment prediction, we also integrate the time series predictive components. We note that these components reduce the variance introduced by noisy observations in the control market time series. Compared to the difference-in-difference based approach deployed by Blake *et al.* in [4], we integrate time series components and rely on the control predictive power of the treatment observations. Brodersen *et al.* provide evidence where state-space methods are more accurate than difference-in-difference techniques to measure the effect of interventions [5].

2.2 Controlled Experiment Design

We identify control and treatment groups based on market-level weekly conversions and market-pair matching combinations. For a given market pair matching, we find the treatment effect without intervention (placebo tests), as described in section 2.1.

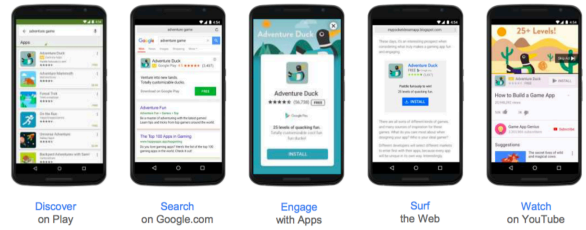


Figure 1: Universal App Campaigns (UAC) media formats within the Ad Network ecosystem. These creative designs are shown for illustrative purposes and do not represent the actual creatives used in the current study. Image source: How to Reach Your App Promotion KPIs.

Algorithm 1 illustrates the process to determine the best control market for a given treatment market. For a given set of markets Ω, a set of possible intervention times Ω, a length of historical data for model fitting Δ_t, and a potential experiment duration after intervention Δ_{t_i}, we run placebo tests (Algorithm 1 lines 8-11).

We consider individual combinations with tighter credible intervals around zero effect to be superior candidates since they diminish the probability of false positives in the hypothesis testing of intervention effects. We keep a subset with the tightest credible intervals that include zero (Algorithm 1 lines 14-15). We filter the best matches where we can maximize our control of advertising spend. The final match selection is chosen before the controlled experiment begins and represents ideal markets based on conversion and spend customization.

Compare to other geo-experiment tests, where the control markets are identified based on correlation [4], we set up the effect estimation method first, and use it to determine the control markets. We note that using the same causal estimation methodology in the market pair selection improves the robustness of the controlled experiment design.

3 MEASURING UAC INCREMENTALITY

3.1 Universal App Campaigns

Universal App Campaigns (UAC) unify ad network's traffic inventory, across the Search Network, Display Network, Youtube, App Store, and Mail. When launching a campaign, the advertiser supplies copy and creative that are reformatted across these various inventory sources. Figure 1 depicts the typical media types within UAC and some ad formats.

UAC claims to have more diverse data points than traditional media channels as a result of its heterogeneity of media types. Due to the access to the app store, UAC campaigns are updated automatically when a given user has installed the advertiser's app. This data advantage makes the audience suppression significantly more efficient for UAC than for other types of advertising.

Despite all the data leverage and the media-type diversity of the UAC ecosystem, advertisers have few customization levers. Even if aggregate user-level ad exposures were provided, non-exposed user populations would be needed to approximate the counterfactual response (see [9] for a survey of observational studies with limited success of measuring incrementality). Therefore, user-level randomization is not possible for testing UAC. Given the increasing popularity of UAC as a different advertising media, evaluating the incrementality of this channel becomes increasingly relevant to the industry.

3.2 Treatment Intervention

To estimate the effectiveness of the UAC media channel, we design the experiment to cut the channel spend in the treatment markets from the ongoing regular advertising strategy. UAC provides the ability to target spend at the DMA level and zip code level. For the current study, we use Uber-cities market design based on the zip codes each one covers. Therefore, we, as an advertiser, have control over weekly spend based on the geo-targeting capabilities of the channel.

Assuming diminishing returns of additional advertising spend [7], we determine that we have a more substantial probability of measuring the channel effects by cutting the spend than by increasing it. This strategy gives us a direct estimation of the incremental conversions generated by the channel, similar to the experiment run by Blake *et al.* for paid search [4].

Due to the UAC mixture of demand-generating and demand-capture media types, changes in UAC spend are likely to have an effect on other channels in the firm advertising media mixture. These spillover effects occurred because there is a significant overlap of audiences among media channels. Thus, other channels' spend is likely to increase to cover the gap of the spend cut of the focal channel, *i.e.* UAC, decreasing the power of the test and under-estimating the value of the channel. Therefore, to avoid spillover effects from other channels, we control the other channels spend to keep the same levels, *i.e.* constant cost-per-attributed-signup (CPA), during the stabilization period in both market groups. We also carefully control the UAC channel spend in both market groups by setting a similar bidding and intra-channel allocation strategy. These potential confounding factors are overlooked in previously proposed geo-experiment designs [4, 13].

We note that the overall controlled experiment design is resource-consuming and potentially expensive. Controlling for other channels' spend to diminish the risk of spillovers requires pausing the complete advertising strategy for the treatment and control markets. That is, spend optimization should be suspended to maintain similar pre-experiment spend levels in both markets. Also, the intervention needs to be significantly large to be detectable in the aggregate conversion (signups) time series. Thus, the spend cut intervention needs to be performed in some of the top-spending markets for the firm. That is likely to represent a sizable strategic price of the test. Despite the cost and effort, this design maximizes the likelihood of successfully measuring the effect of the media channel. Even when we expect an incremental effect of the channel a priori, in practice, proving channel incrementality rigorous estimates is challenging (see [9, 15] for empirical evidence of the difficulties). As a result, special care needs to be taken in the intervention design.

4 SETUP AND VALIDATION

To identify the best market matchings, we run placebo tests (Algorithm 1) for a sample of 50 most representative Uber markets (Ω), leading to 2,450 market matching combinations. We consider a set of different intervention dates and training periods to decrease

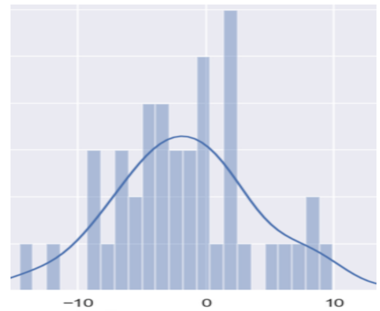


Figure 2: Placebo effect distribution to select best control markets. x-axis represents lift (%) effect point estimates.

the effect of arbitrary choices of an intervention date for placebo treatments. We test for 5 intervention times (Φ): winter, spring, summer, fall, and holidays. We consider 6 months of historical data prior to the intervention date for training (Δ_t) and 2 months of experiment duration after the intervention (Δ_{t_i}). Figure 2 shows the distribution of lift effect point estimates of Algorithm 1 lines 8-11. We observe a large variability of effects, including values different from zero. This process illustrates the need for a more sophisticated market matching than a simple correlation analysis (as proposed by [4]).

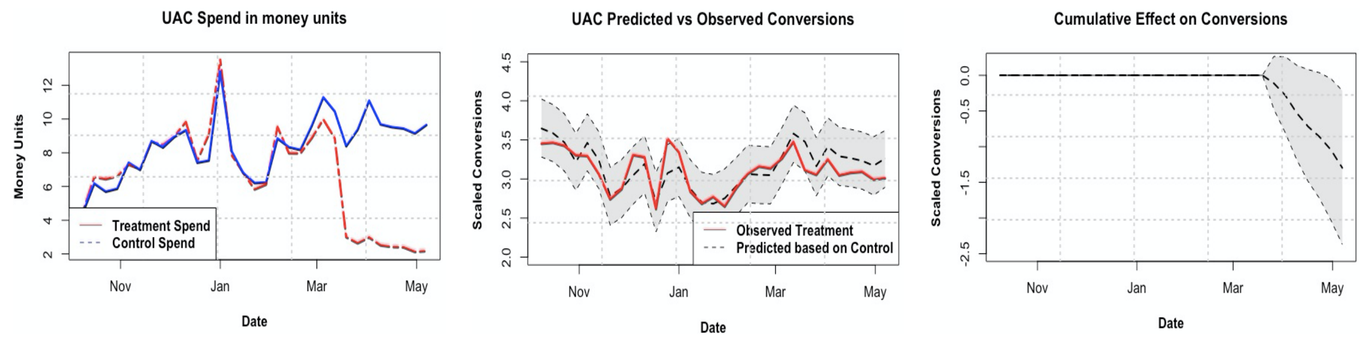
We identify that holidays are often problematic as conversions behave differently for each market. These different behaviors increase the variability of our estimates, making it more challenging to measure the effectiveness of the channel accurately. To avoid the impact of holidays, we execute the experiment intervention on March 19th, 2018, after stabilizing the channel spend in both cells starting on January 15th, 2018. The experiment ended on May 15th, 2018. Figure 3(a) shows the historical spend for both markets, the spend stabilization period, and the spend cut at the time of intervention.

Based on Algorithm 1 lines 14-15, we optimize the market pair selection to minimize the probability of a false positive effect detection. For the current experiment, we find a minimum detectable lift of 0.83% based on the best control for the selected treatment market, and the settings we set above.

5 RESULTS

Figure 3 shows the experiment spend intervention, the treatment conversions against the predictive control conversions, and the cumulative effect of the treatment, based on the metric definitions of Equation 3.

We observe in Figure 3(b) noisier observations before the spend normalization period. Here, the treatment scaled conversions show slight differences with the predicted scaled conversions without intervention. Reasons for these variations include different historical spend from other channels and the Christmas holidays. Testing based purely on synthetic control solutions need to handle this variability and increase of variance in the effect estimation [1]. We note that during the spend stabilization (01/15/18 - 03/12/18) observed and predicted treatment scaled conversions are closely aligned. This stabilization is one of the primary benefits of our controlled experiment design because it decreases the variance of the effect estimations improving the power of the experiment design.



(a) (b) (c)

Figure 3: Treatment and control weekly time series for: (a) UAC spend, $x_t^{(control)T}$. (b) UAC treatment scaled signups, $y_t^{(treat)}$, and control predictive scaled signups, $\hat{y^s}_t^{(treat)}$, with 95% credible intervals. (c) Cumulative estimated impact, $y^s_{cumAttr(t_i)}$, with 95% credible intervals. Training period, $t \in 1, \ldots, T-1$: 09/09/2017 - 03/12/208. Spend stabilizing period: 01/15/18 - 03/12/18. Spend cut intervention date, T: 03/19/18. Ending date, T_i: 05/13/18. Test duration: 8 weeks. MCMC samples $N_s = 15,000$.

Table 1: UAC Campaign Controlled Experiment Results.

Estimation Method	Metric	Treatment Lift	P-Value	Credible Interval Lower Bound	Credible Interval Upper Bound	Incremental Scaled Units	Cost per Incremental Unit
Proposed Estimation Method	Scaled Signups	-6.57%***	0.0064	-11.66%	-1.43%	-1.29	39.30 money units
Synthetic Control	Scaled Signups	-5.03%	0.2180	-13.73%	3.66%	-1.01	50.20 money units
Proposed Estimation Method	Spend Units	-72.33%***	0.0010	-82.40%	-61.28%	-50.70	-

As depicted in Figure 3(b), we observe a consistent drop in the treatment scaled conversions after the intervention during the experiment when compared to the predicted scaled conversions. The drop is within the credible intervals of the predicted treatment scaled conversions on a week-by-week basis. Similar to the case of display advertising effectiveness measurement [15], detecting statistically significant effects due to UAC advertising spend is difficult. Despite the demand-capturing components in the UAC intra-channel mix (*i.e.* sponsored search, app store), and the drastic intervention by cutting the spend close to zero, measuring the effectiveness of the channel remains challenging. Figure 3(c) shows the cumulative effect of the spend cut, which becomes significant in the last two weeks of the experiment. We observe an increasingly significant impact, which is the result of a consistently negative effect on the week-by-week point estimates.

We report the cumulative effect lift results, $lift^s_{cum(t_i)}$, after concluding the experiment in Table 1. Based on this cumulative result, we conclude that the channel is incremental and that UAC spend would have caused 6.57% of the scaled conversions that would have been observed in the average. We observe in Figure 3(a) that the spend is not completely zero. That is because of tailing ad exposures that require spend delivery to the ad network during the experiment. Therefore, to find the cost per incremental scaled conversion, we estimate the effect of the intervention on spend. We report an effect on spend of -72.33% in Table 1. As a result, the cost per incremental scaled conversion becomes the ratio of incremental spend over incremental scaled conversions (CPIA). We report a CPIA of 39.30 money units, which reflects the return on investment of UAC advertising spend in the current experiment.

For comparison, we use an internal synthetic-control based method off-the-shelf [3]. This method uses a standard synthetic control approach [1] based on the control market conversions used by the proposed method. Based on the results of Table 1, we observe that this synthetic-control based estimation provides a similar point estimate but with larger p-value, without achieving statistical significance. We attribute the more substantial variability to the lack of spend stabilization and the dependence on historical spend in the predictive markets, despite the usage of more conversion time series predictors. This increase in the precision of the results shows the value of spend stabilization to successfully measure the effect of the campaign with statistically significant results.

6 LIMITATIONS

Accurately measuring the effect of advertising on a given conversion metric is critical for successful advertising spend planning. Identifying these effects is hard, and often leads to a large percentage of inconclusive campaign experiments [15]. Comparisons between aggregate market conversions require large intervention effects. Without user-level data, we are unable to identify the user population exposed to the campaign ads and isolate their conversions. The smaller the exposed population of a given market, the more diluted the spend effect is. Adding any filter on the users who we are confident are not exposed to ads improves the likelihood of successfully measuring a significant effect. As a result, the proposed design and estimation framework require interventions with large expected effects. Since the incremental value of additional advertising spend over ongoing spend levels is smaller, assuming

[3] *Under the Hood of Uber's Experimentation Platform.* Published: https://eng.uber.com/xp

diminishing returns of incremental budget [7], we recommend suspending spend instead.

Holidays and unexpected events are problematic to model in market-based experiments. The seasonal effects of holidays are often confounded with market intrinsic properties, which make them difficult to predict their conversions. Consequently, the experiment precision decreases, increasing the likelihood of false positives in the hypothesis testing of effects. The general recommendation is to discard and assume holidays as missing values [13]. In the current experiment, we decide the experiment dates in the market match procedure based on thousands of placebo tests.

The current controlled experiment design is not orthogonal to other channel's spend in any of the treatment or control markets. Given the typical overlap in user audiences among channels, pausing all advertising spending in the treatment market creates a marketing gap that is easily picked by other channels. Thus, the advertising optimization strategy needs to be controlled in both treatment and control markets during the spend stabilization period. There are cases where a market match test is preferred over a randomized A/B experiment. Cases when negative brand perceptions are likely to spread if users without access to the advertising offer become aware that they have been excluded. We recommend randomized A/B experiments with user-level exposures and responses whenever possible. Unfortunately, often ad networks do not provide this capability, and UAC exemplifies these constraints.

7 CONCLUSION AND MANAGERIAL IMPLICATIONS

We have presented a controlled experiment design and effect estimation framework with advertiser side spend levers when user level randomization is not possible. We have detailed and discussed the day-to-day constraints faced by advertisers to effectively measure advertising channel incrementality with experiments. These constraints contrast with the majority of the current literature in ad effectiveness which assumes the ability to randomize users.

We approached the incrementality testing problem with a market matched based experiment and by targeting advertising spend at the market level. We have provided practical recommendations to successfully measure channel incrementality with advertiser spend levers without user level exposure data. We have demonstrated that the proposed method is more effective in detecting and estimating the effect of channel spend than standard synthetic control based approaches.

We have provided evidence from a large-scale field experiment that UAC, a recent online advertising format, shows incremental value as a media channel, despite the limited customization levers. We hypothesize that the diversity of the media type within UAC is one of the main reasons the channel is effective. The combination of demand-generating and demand-capturing advertising media suggests a powerful set of advertising levers. Also, the use of app store within UAC is a powerful tool to automatically blacklisted users who have been acquired, when compared to other media types. As future work, we encourage more testing of UAC marketing campaigns. Measuring the effects of this channel in other types of conversions, apart from customer acquisition, would shed light on the effectiveness of UAC more broadly.

Testing without user-level randomization is expensive, but it is the only option without user randomization. Although some ad networks provide incrementality tools, particularly in ad exchanges and programmatic display, the fragmented nature of the online advertising industry still poses significant constraints. In ongoing advertising evaluation and optimization planning, rigorously designed experiments provide valuable data to build channel response curves for incremental conversions and to Media Mix Models. We recommend running experiments to calibrate those models. Making rigorous experimentation for advertisers more effective and across multiple media channels remains an open research topic.

ACKNOWLEDGMENTS

We thank Katie He, who played a key role in the stabilizing of the spend efficiency, and to Kevin Neifach the UAC channel manager. We thank Tina Nikou for the help in earlier versions of the paper.

REFERENCES

[1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association* 105, 490 (2010), 493–505.

[2] Joel Barajas, Ram Akella, Marius Holtan, and Aaron Flores. 2016. Experimental designs and estimation for online display advertising attribution in marketplaces. *Marketing Science* 35, 3 (2016), 465–483.

[3] Joel Barajas, Ram Akella, Marius Holtan, Jaimie Kwon, Aaron Flores, and Victor Andrei. 2012. Dynamic Effects of Ad Impressions on Commercial Actions in Display Advertising. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1747–1751. https://doi.org/10.1145/2396761.2398510

[4] Thomas Blake, Chris Nosko, and Steven Tadelis. 2015. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* 83, 1 (2015), 155–174.

[5] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2015. Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* 9, 1 (03 2015), 247–274. https://doi.org/10.1214/14-AOAS788

[6] Kay H. Brodersen and Alain Hauser. 2017. CausalImpact. An R package for causal inference in time series. https://google.github.io/CausalImpact/

[7] David Chan and Mike Perry. 2017. *Challenges and Opportunities in Media Mix Modeling*. Technical Report. Google.

[8] Facebook. 2019. About Facebook Conversion Lift Studies. https://www.facebook.com/business/help/688346554927374

[9] Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2018. A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Forthcoming at Marketing Science* (2018).

[10] Hemant Ishwaran and J. Sunil Rao. 2005. Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* 33, 2 (04 2005), 730–773. https://doi.org/10.1214/009053604000001147

[11] Garrett A. Johnson, Randall A. Lewis, and Elmar I. Nubbemeyer. 2017. Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness. *Journal of Marketing Research* 54, 6 (2017), 867–884. https://doi.org/10.1509/jmr.15.0297 arXiv:https://doi.org/10.1509/jmr.15.0297

[12] Kirthi Kalyanam, John McAteer, Jonathan Marek, James Hodges, and Lifeng Lin. 2018. Cross channel effects of search engine advertising on brick & mortar retail sales: Meta analysis of large scale field experiments on Google.com. *Quantitative Marketing and Economics* 16, 1 (01 Mar 2018), 1–42. https://doi.org/10.1007/s11129-017-9188-7

[13] Jouni Kerman, Peng Wang, and Jon Vaver. 2017. *Estimating Ad Effectiveness using Geo Experiments in a Time-Based Regression Framework*. Technical Report. Google.

[14] Pavel Kireyev, Koen Pauwels, and Sunil Gupta. 2016. Do display ads influence search? Attribution and dynamics in online advertising. *International Journal of Research in Marketing* 33, 3 (2016), 475–490.

[15] Randall A. Lewis and Justin M. Rao. 2015. The Unfavorable Economics of Measuring the Returns to Advertising *. *The Quarterly Journal of Economics* 130, 4 (2015), 1941–1973.

[16] Pengyuan Wang, Wei Sun, Dawei Yin, Jian Yang, and Yi Chang. 2015. Robust Tree-based Causal Inference for Complex Ad Effectiveness Analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. ACM, New York, NY, USA, 67–76. https://doi.org/10.1145/2684822.2685294