A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook*

Brett Gordon
Kellogg School of Management
Northwestern University

Florian Zettelmeyer
Kellogg School of Management
Northwestern University and NBER

Neha Bhargava
Facebook

Dan Chapsky
Facebook

June 27, 2016

PRELIMINARY AND INCOMPLETE
DO NOT CITE

Abstract

Advertisers are keenly interested in knowing the effectiveness of their online advertising. However, the industry seldom uses randomized experiments to estimate effectiveness, relying instead on observational methods such as matching and regression. This is partly because, until recently, randomized experiments have been difficult or expensive to implement in online advertising contexts, and partly because observational methods are widely considered within the industry to be "good enough." We analyze whether observational methods for causal inference can reliably substitute for randomized experiments in online advertising measurement. This is of particular interest because there have been enormous recent improvements in observational methods for causal inference (Imbens and Rubin 2015). Using data from 12 US advertising lift studies at Facebook comprising 435 million user-study observations and 1.4 billion total impressions, we contrast the experimental results to those obtained from a variety of observational methods. We show that observational methods often fail to produce the same results as the randomized experiments, even after conditioning on information from thousands of behavioral variables and using non-linear models. Our findings suggest that common approaches in industry used to measure advertising effectiveness fail to measure accurately the true effect of ads.

* No data contained personally identifiable information that could identify consumers or advertisers to maintain privacy. We thank Daniel Slotwiner, Gabrielle Gibbs, Joseph Davin, Brian d'Alessandro, and Fangfang Tan at Facebook and seminar participants at CKGSB, Columbia, ESMT, HBS, Northwestern, and Temple for helpful comments and suggestions. We particularly thank Meghan Busse for extensive comments and editing suggestions. Gordon and Zettelmeyer have no financial interest in Facebook and were not compensated in any way by Facebook or its affiliated companies for engaging in this research. E-mail addresses for correspondence: b-gordon@kellogg.northwestern.edu, f-zettelmeyer@kellogg.northwestern.edu, nehab@fb.com, chapsky@fb.com

1 Introduction

Digital advertising spending is expected to exceed television advertising spending for the first time in 2016. Firms are projected to spend $160 billion worldwide and $66 billion in the US on digital advertising.[1] Not surprisingly, advertising is one of the key funding sources for Internet content and services. For example, the five most visited sites in the world—Google, Facebook, Youtube, Baidu, and Yahoo—all rely on advertising revenues as their business model.[2]

As advertisers have shifted a larger fraction of their ad expenditures online, so has demand grown for online ad effectiveness measurement. As a result, advertisers now routinely demand and have access to granular-level data that link ad exposures, clicks, page visits, online purchases, and in some cases offline purchases. When these measures are used to evaluate an advertising campaign with a randomized experiment, or randomized controlled trial (RCT), advertisers generally have the necessary data to estimate the causal effect of the ad campaign.

In practice, however, few online ad campaigns are evaluated using RCTs (Lavrakas 2010). This is for multiple reasons. First, technical limitations of advertising platforms often make experimentation cumbersome and labor intensive. Even large online advertisers such as Google and Facebook enabled clients only in 2015 to perform advertising RCTs at the cookie or individual level. Second, advertising RCTs can be seen as expensive. To see why, note that online RCTs have traditionally been run using Public Service Announcements (PSAs) as control ads (Johnson, Lewis, and Nubbemeyer 2015b). If so, the advertiser has to pay for PSAs, therefore funding ads that will have no benefit for the firm. Third—and most importantly—RCTs are seen by many in the industry as unnecessary in light of observational methods that rely on comparing consumers who were exposed to an ad with appropriately chosen consumers who were not (Gluck 2011). In line with this, leading ad measurement companies rely on such observational methods to estimate the causal effect of advertising (Abraham 2008, comScore 2010, Klein and Wood 2013).

However, due to selection of which consumers are exposed to ads, estimating the causal effect of ads based on observational methods is far from straightforward. Selection arises, first, because advertising platforms try to serve ads to consumers who are more likely to buy. Second, selection arises because consumers have to be online to be exposed to an online ad. If the success metric for the ad campaign is purely online—an online purchase, a registration, etc.—exposed users will be more likely to convert simply because they happened to be online during the campaign. Lewis, Rao, and Reiley (2011) show that this *activity bias* complicates measuring causal effects online.

[1] `http://www.nytimes.com/2015/12/07/business/media/digital-ad-spending-expected-to-soon-surpass-tv.html`, accessed on 3-23-2016.

[2] `http://www.alexa.com/topsites`, accessed on 3-23-2016.

In this paper we analyze whether and when observational methods can reliably substitute for randomized experiments in online advertising measurement. We do so by using a collection of 12 large-scale advertising RCTs conducted at Facebook. We use the outcomes of these studies to reconstruct different sets of observational methods for measuring ad effectiveness and then compare each of them to the results obtained from the RCT.

Our data allows us to avoid common challenges faced in online advertising measurement. Most advertising data is collected at the level of a web browser cookie. While measurement firms try to determine which cookies belong the same individual, this is not always possible. This has two potential consequences. First, users in an experimental control group may inadvertently also be simultaneously assigned to the treatment group. Second, advertising exposure across devices may not be fully captured. We avoid both of these problem because Facebook requires users to log into Facebook each time they access the service on any browser and device. This means that ads are never show inadvertently to users in the control group and all ad exposures are measured. In addition, our ad campaigns were selected to have measures of purchase outcomes in addition to registrations and web page views.

We find that across the advertising studies, on average, there is a significant discrepancy between the observational approaches and RCTs. To illustrate, the advertising campaigns we analyzed caused (based on the RCT), on average, a 57% increase in purchases by Facebook users (in the industry this is referred to as the "lift" of the ad campaign). The observational methods we analyzed yielded lift estimates that were (in absolute terms) between 173 and 661 percentage points different from the 57% RCT estimate.

In addition, we find that there is tremendous variation in how close the observational methods we analyzed came to replicating the RCT estimates. In some studies, observational methods that used the most detailed data yielded estimates that were statistically indistinguishable from RCT estimates. In other studies observational estimates were massively higher than those from the RCT. Finally, in some studies the observational estimates were lower than those from the RCT. Based on the limited set of studies at our disposal we could not identify characteristics of ad campaigns that would lead to one or the other outcome.

Our paper makes two contribution. First, we shed light on whether—as is commonly believed in the industry—observational methods for ad measurement are "good enough." Answering this question is important in the light of an ongoing debate on the role of RCTs (see, for example, Deaton (2010) and Hausman (2016)). Applied to the context of advertising, the critique is that RCTs can only test relatively few advertising strategies relative to the enormous possible variation in advertising; which creative? which publisher? which touchpoint? which sequence? which

frequency? This highlights one potential benefit of observational methods, which is that, relative to RCT's, much more data for high-dimensional problems is typically available because the data are generated more easily and by more actors. The usual issues of selection bias, suitable controls, etc., must be addressed, and of course this is exactly what observational methods try to do. Our paper sheds some light on whether these efforts succeed in the context of online advertising.

Second, we contribute to a literature on observational vs. experimental approaches to causal measurement. In his seminal paper, Lalonde (1986) compares observational methods with randomized experiments in the context of the economic benefits of employment and training programs. He examined the "... results likely to be reported by an econometrician using non experimental data and the most modern technique ..." (p. 604) with those of a field experiment. He concluded that "... many of the econometric procedures do not replicate the experimentally determined results ..." (p. 604). Since then, we have seen enormous improvements in observational methods for causal inference (Imbens and Rubin 2015). In fact, Imbens (2015) shows that an application of these improved methods to the Lalonde (1986) dataset manages to replicate the experimentally determined results. In this paper we analyze whether the improvements in observational methods for causal inference are sufficient for replicating experimentally generated results in a large industry where such methods are commonly used in practice.

We are far from the first to write about online advertising effectiveness.[3] In one of the first papers on the subject, Lewis and Reiley (2014) run a field experiment in which they link advertising to retail sales. The paper shows that sample sizes have to be very large in order to measure the effect of online ads on purchase outcomes. Building on this work, Lewis and Rao (forthcoming) use twenty-five large advertising field studies to quantify how the volatility of consumer expenditure affects power in measuring ad effects. Lewis, Rao, and Reiley (2011) are the first to identify and document the unobserved characteristics that lead to "activity bias," which we have described above. This paper is also one of the earliest works to compare advertising effect estimates from a RCT to those obtained using observational methods (exposed vs. unexposed and regression). In the best case, they find a regression overstates the RCT causal effect by a factor of 161.[4]

A variety of related work on advertising effects exists. Lewis and Nguyen (2015) show that complementarities across display and search advertising are very hard to reliably estimate (see also Rutz and Rucklin (2011)). Others have used purchase intent surveys to show the value of targeting

[3]For a comprehensive summary of issues addressed in this literature please see Lewis, Rao, and Reiley (2015).

[4]One challenge in Lewis, Rao, and Reiley (2011) was finding a good pseudo-control group of unexposed users for the observational methods. The reason is that the experiment exposed 95 percent of US-based traffic to the focal ad and held out 5 percent as the experimental control. Consequently, the pseudo-control group consisted entirely of international users who visited www.yahoo.com on the day of the experiment but were not targeted by the ad.

ads and the effect of context and intrusiveness of ads (Goldfarb and Tucker 2011a, Goldfarb and Tucker 2011b). Johnson, Lewis, and Nubbemeyer (2015b) introduce a method for implementing RCTs for display advertising. Johnson, Lewis, and Reiley (2015) measure the effect of repeated exposure and proximity to the advertiser. Johnson, Lewis, and Nubbemeyer (2015a) show the distribution of ad effectiveness across 431 ad experiments from varied industries. Sahni and Nair (2016) use a series of mobile experiments to tease apart the mechanism underlying native advertising on a restaurant search platform. A closely related paper to our own is Blake, Nosko, and Tadelis (2015) which documents that non-experimental measurement can lead to highly suboptimal spending decision for online search ads. However, in contrast to our paper, Blake, Nosko, and Tadelis (2015) is based on randomization at the level of major markets. In summary, while there is a rich recent literature on online advertising, we are among the first to be able to evaluate observational methods for causal inference using the individual-level data advertisers can observe.

This paper proceeds as follows. We begin by describing in the next section the experimental design used in the 12 advertising RCTs we analyze. We describe how advertising works on Facebook, how RCTs are implemented, and what determines advertising exposure. In Section 3 we introduce the potential outcomes notation that has become standard for causal inference and relate it to analyzing our RCT. In Section 4 we introduce the observational methods we use. Section 5 introduces the data and shows randomization checks. Next, we report on the results; for expositional reasons we begin in Section 6 by showing all results for one example ad campaign in detail. Section 7 then summarizes the findings for all remaining ad campaigns. Section 8 offers some concluding remarks.

2 Experimental Design

This section provides a description of how Facebook conducts advertising campaign experiments.[5] First, we provide some relevant background information about the Facebook advertising platform. Second, we define the measurement question of interest to the advertiser and relate it to the experimental implementation. Third, we discuss the determinants of advertising exposure for users assigned to the test group, where compliance is one-sided.

2.1 Advertising on Facebook

We focus exclusively on campaigns where the advertiser designed the campaign with a particular "direct response" outcome in mind, for example to increase sales of a new product, to attract new

[5]Within Facebook these ad tests are referred to as "lift tests." See `https://www.facebook.com/business/news/conversion-lift-measurement`.

customers, etc.[6] Such outcomes are referred to as a "conversion outcome" in the industry. In each study the advertiser measured outcomes using a piece of Facebook-provided HTML code, referred to as a "conversion pixel," that the advertiser embeds on its web pages.[7] This enables an advertiser to measure whether a user visited that page. These pixels can be placed on a variety of pages and therefore measure different conversion outcomes. For example, if the conversion pixel is placed on a checkout confirmation page the pixel measures a purchase outcome. If the conversion pixel is placed on a registration confirmation page, the pixel measures a registration outcome, etc. These pixels allow the advertiser (and Facebook) to record conversions irrespective of whether the user was in the control or test group and does not require the user to click on the ad to have her conversion outcomes measured.

Facebook's ability to track users via "single-user login" across devices and sessions represents a significant measurement advantage over common cookie-based settings. First, this helps to ensure the integrity of the random assignment mechanism because a user's assignment can be maintained persistently throughout the campaign and prevents control users from being inadvertently shown an ad. Second, Facebook can associate all exposures and conversions, across devices and sessions, with a specific user. This is important because users are frequently exposed to advertising on a mobile device but might subsequently convert on a tablet or computer.

Figure 1 displays where a Facebook user accessing the site from a desktop/laptop or mobile device might see ads. In the middle is the "News Feed," where new stories appear with content as the user scrolls down the page or the site automatically refreshes. Ad impressions on Facebook are served in the News Feed interlaced with organic content, with a smaller portion served to the right of the page. On mobile devices, only the News Feed is visible and so no ads appear on the right side.[8]

In the News Feed, an ad impression is represented by one tile. As the user scrolls and new tiles appear, new ad impressions are interspersed with regular content. The rate at which Facebook serves ads in the News Feed is carefully managed at the site-level to maintain the proper user experience.

[6] This excludes brand building campaigns where outcomes are measured through consumer surveys.

[7] We use "conversion pixel" to refer to two different types of conversion pixels used by Facebook. One was traditionally referred to as a "conversion pixel" and the other is referred to as a "Facebook pixel". Both types of pixels were used in the studies analyzed in this paper. For our purposes both pixels work the same way (see `https://www.facebook.com/business/help/460491677335370`).

[8] News Feed ads are an example of "native advertising" because the ads are interlaced with the page's organic content. This style of ads is increasingly common because they are thought to interfere less with the user experience and still potentially to have better response rates. Sahni and Nair (2016) use a series of mobile experiments to tease apart the mechanism underlying native advertising on a restaurant search platform.

Figure 1: Facebook desktop and mobile ad placement

https://www.facebook.com/business/ads-guide

An advertising campaign is a collection of related advertisements (creatives) served during the campaign period. A campaign may have multiple creatives associated with it, as Figure **??** illustrates for Jasper's Market, a fictitious advertiser commonly used in examples at Facebook. Although the imagery and text vary from one creative to the next, the overall message in the campaigns we reviewed were generally consistent. Our analysis focuses on evaluating the effect of the campaign as a whole rather than on the effects of specific ad creatives.

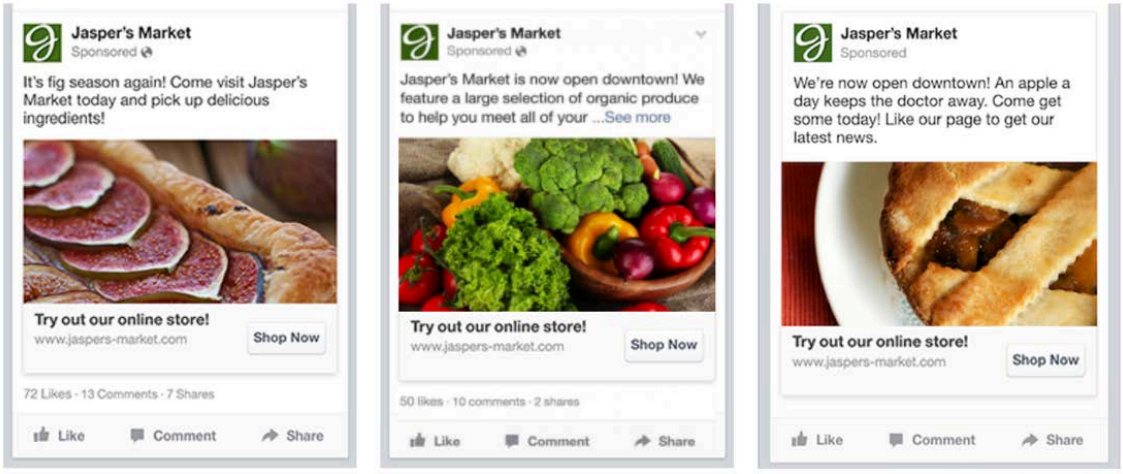
As with most online advertising, each impression is the result of an underlying auction. The *opportunity set* is the collection of display ads that are entered into the auction to bid for an impression. The advertising platform determines which ads are part of the opportunity set based on a combination of factors such as: how recently the user was served any ad in the campaign, how recently the user saw ads from the same advertiser, the overall number of ads the user was served recently, whether the user is in the target audience as defined by the advertiser, etc.[9] The auction plays a role both in the implementation of the experiment and in generating endogenous variation in exposures.

2.2 Experimental Implementation

An experiment begins with the advertiser defining a new marketing campaign which includes deciding which consumers to target. For example, the advertiser might want to reach all users that

[9]The relevance score attempts to adjust for whether a user is likely to be a good match for an ad (`https://www.facebook.com/business/news/relevance-score`).

Figure 2: Example of three display ads for one campaign



match a certain set of demographic variables, e.g., all women between the ages of 18 and 54. This choice determines the set of users included in the study sample. Each user in the study sample is randomly assigned to either the control group or the test group according to some proportion selected by the advertiser (in consultation with Facebook). Users in the control group are never exposed to campaign ads during the study and users in the test group are eligible to see the campaign's ads during the study. Whether these eligible users end up being exposed to the ads depends on a variety of factors, for example, whether the user accessed Facebook during the study period. We discuss these factors and their implications in detail in the next subsection. As a consequence, we observe three groups of users: control-unexposed, test-unexposed, and test-exposed.

Next we consider a critical question: for users in the control group, what ads should they be shown in place of the advertiser's campaign? This question defines the counterfactual of interest. To evaluate campaign effectiveness, an advertiser requires the control condition to estimate the outcomes that would have occurred in the absence of the campaign. Thus, the ads served in the control condition in place of the focal campaign should be the ads that *would have been* served if the advertiser's campaign had not taken place.

We illustrate how this process works using a hypothetical and stylized example in Figure 3. Consider two users in the test and control groups, respectively. Suppose that at one particular instant, Jasper's Market wins the auction to display an impression for the test group user, as seen in Figure 3a. Imagine that the control group user, who occupies a parallel world to the test user, would have been served the same ad had this user been in the test group. However, the platform, recognizing the user's assignment to the control group, prevents the focal ad from being displayed. As Figure 3b shows, instead the second-place ad in the auction is served to the control user because

it is the ad that would have won the auction in the absence of the focal campaign.

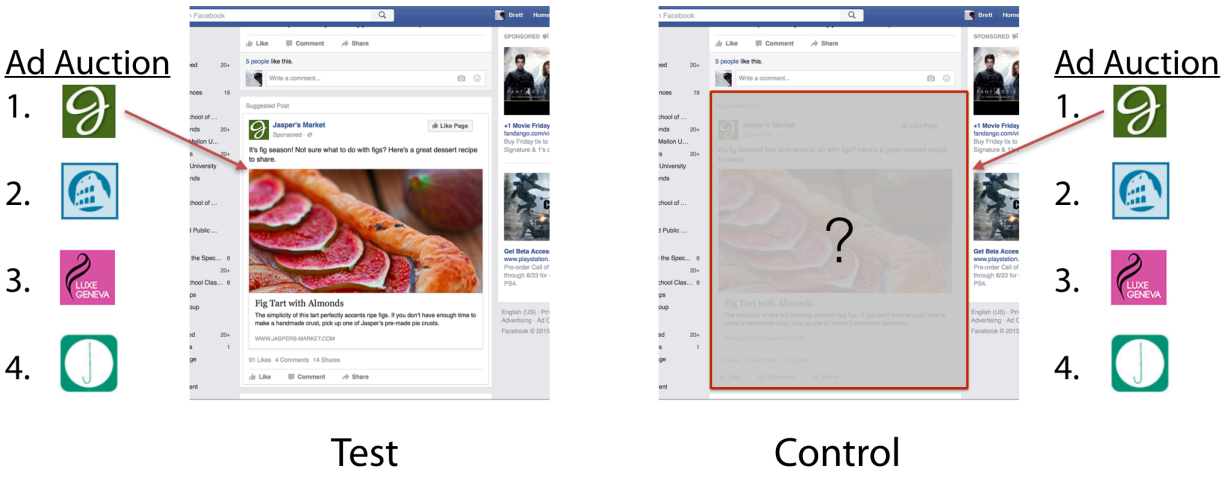
It is worth emphasizing a few points about this mechanism. Figure 3 is a stylized view of the experimental design (clearly there are no two identical users in parallel worlds). For users in the test group, the platform serves ads in a regular manner, including those from the focal advertiser. The experimental mechanism is only relevant for users in the control group and if the opportunity set contains the focal ad on a given impression. If the focal ad does not win the auction, there is no intervention—whatever ad wins the auction is served because the same ad would have been served in the absence of the focal advertiser's campaign. However, if the focal ad wins the auction, the system removes it and instead displays the second-place ad. In the example, Waterford Lux Resorts is the "control ad" shown to the control user. At another instant when Jasper's Market would have won the auction, a different advertiser might occupy the second-place rank in the auction. Thus, instead of their being a single control ad, users in the control condition are shown the distribution of ads they would have seen if the advertiser's campaign had not run.

This approach relies on the fact that the auction mechanism is stable to the removal of the focal ad. That is, the second-place ad is the same regardless of whether the focal advertiser participated or not in the auction.[10] This assumes other advertisers' strategies are fixed in the short-run and do not respond to the fact that the focal advertiser is running the campaign. This assumption is reasonable because campaigns are not pre-announced, occur over relatively short periods, and it would be hard for another advertiser to gauge the scope of any campaign given the scale of Facebook. In addition, there are a wide variety of advertisers, each with different targeting specifications. As a result, the chance that the ads of two competitors are the highest and second-highest ad are very low.

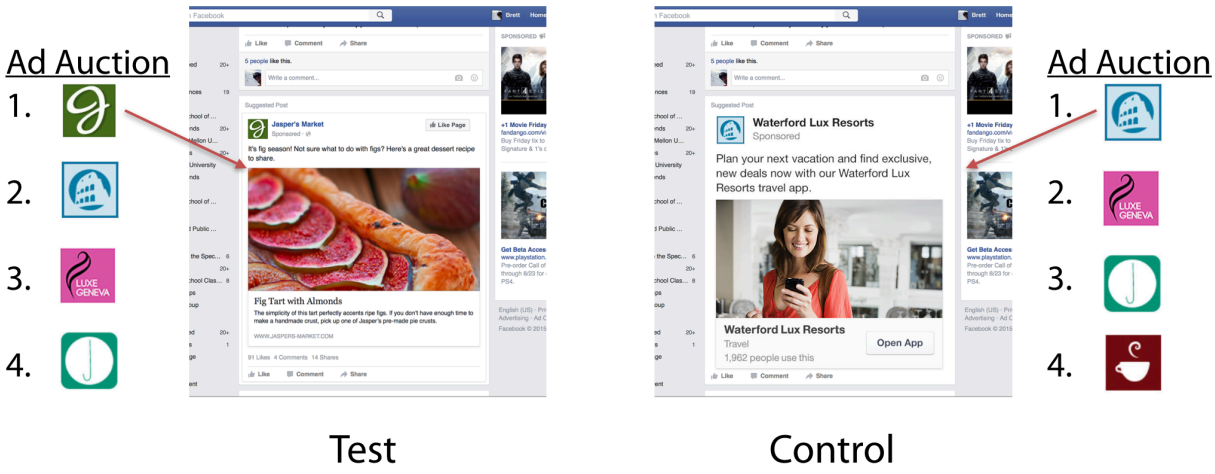
Finally, as with any experiment, this one yields an estimate of the average treatment effect of the campaign conditional on all market conditions. This includes any marketing activities the advertiser is conducting in other channels (e.g., search, TV) and the activities of its competitors. The estimated lift obtained during the experiment may not generalize to similar future campaigns if market conditions change. If advertising effects are nonlinear across mediums, which is possible, the experiment measures something akin to the average *net* effect of the campaign given the distribution of non-Facebook advertising exposures across the sample.

[10]The auction is a modified version of a second-price auction such that the winning bidder pays only the minimum amount necessary to have won the auction. However, due to the additional factors considered in the bid ranking (e.g., the relevance score), bidders may not be ranked strictly based on their bid prices. For more information, see `https://www.facebook.com/business/help/430291176997542`.

Figure 3: Determination of control ads in Facebook experiments



(a) Step 1: Determine that a user in the control would have been served the focal ad



(b) Step 2: Serve the next ad in the auction.

2.3 Determinants of Advertising Exposure

In our setting, compliance is perfect for users in the control group, who are never shown any campaign ads. However, compliance is only one-sided in the test group, where exposure (receipt of treatment) is an endogenous outcome that depends on factors such as the user, advertisers, and the platform. These factors generate systematic differences (i.e., selection bias) between the exposed and unexposed users in the test group. Three general features of online advertising environments, not just limited to Facebook, make the selection bias of exposure particularly significant.

First, an ad is delivered when the advertiser wins the underlying auction for an impression. Winning the auction implies the advertiser out-bid the other advertisers competing for the same impression. Additionally, Facebook and some other publishers prefer to show ads to consumers they are more likely to find interesting and useful.[11]. This means that an advertisers' ads are more likely to be shown to users who are more likely to respond to its ads (measured by the relevance score), *and* users who are less likely to respond to the other advertisers who are currently active on Facebook. Even if an advertiser triggers little selection bias based on their own advertising, it can nevertheless end up with a selected exposure because of what another advertiser does. For example, if another advertiser if placing high bids on mothers during the campaign period, there is a higher likelihood that mothers will not be exposed to the campaign.

A second mechanism that drives selection is the optimization algorithms that exist on modern advertising delivery platforms. Advertisers and platforms try to optimize the types of consumers that should be shown an ad. For a campaign that seeks to optimize on purchases, a machine learning algorithm will gradually refine the targeting and delivery rules to identify users who are most likely to convert. For example, suppose an advertiser initially targets female users between the ages of 18 and 55. After the campaign's first day, the platform observes that females between 18 and 34 are especially likely to convert. As a result, the ad platform will increase the frequency that the ad campaign enters into the ad auction for this set of consumers, resulting in more impressions targeted at this narrower group. These optimization routines perpetuate an imbalance between exposed and unexposed test group users: the exposed group will contain more 18-34 females and the unexposed group will contain more 35-55 females. Assessing ad effectiveness by comparing exposed vs. unexposed consumers will therefore overstate the effectiveness of advertising because exposed users were specifically chosen on the basis of their higher conversion rates.[12]

The final mechanism arises from the simple observation that a user must actually visit Facebook

[11] See `https://www.facebook.com/help/562973647153813`

[12] Facebook's ad testing platform is specifically designed to account for the fact that targeting rules for a campaign change over time. This is a accomplished by applying the new targeting rules both to test and control groups, even though users in the control group are never actually exposed to campaign ads.

during the campaign to be exposed. If conversion is purely a digital outcome (e.g., online purchase, registration, key landing page), exposed users will be more likely to convert simply because they happened to be online during the campaign. For example, a user on vacation may be less likely to not only visit Facebook but to also engage in various online activities. Lewis, Rao, and Reiley (2011) term this form of selection as activity bias and show its effects can be substantial.

In the RCT, we address this selection bias by leveraging both the random assignment mechanism and whether a user receives treatment. For the observational models, we discard the randomized control group. These methods must deal with the selection bias relying solely on treatment status and observables in the test group.

3 Analysis of the RCT

We use the potential-outcome notation that has become standard in the literature on both experimental and nonexperimental program evaluation. Our exposition in this section and the next draws heavily on material in Wooldridge (2002), Imbens (2004), Imbens and Wooldridge (2009), and Imbens and Rubin (2015).

3.1 Definitions and Assumptions

Each ad study contains N individuals (units) indexed by $i = 1, \ldots, N$ drawn from an infinite population of interest. Although N varies across studies, we do not index any variable by a study-specific subscript because all of our analysis takes place within a study. Individuals are randomly assigned to test or control conditions through $Z_i = \{0, 1\}$. Exposure to ads is given by the indicator $W_i(Z_i) = \{0, 1\}$. Users assigned to the control condition are never exposed to any ads from the study, $W_i(Z_i = 0) = 0$. However, assignment to the test condition does not guarantee a user is exposed, such that $W_i(Z_i = 1) = \{0, 1\}$ is an endogenous outcome outside of the advertiser's control. We also observe a set of covariates $X_i \in \mathbb{X} \subset \mathbb{R}^K$ for each user that are unaffected by the experiment.

Given an assignment Z_i and a treatment $W_i(Z_i)$, the potential outcomes are $Y_i(Z_i, W_i(Z_i)) = \{0, 1\}$. Under one-sided noncompliance, the observed outcome is:

$$Y_i^{obs} = Y_i(Z_i, W_i^{obs}) = Y_i(Z_i, W_i(Z_i)) = \begin{cases} Y_i(0,0), & \text{if } Z_i = 0, W_i^{obs} = 0 \\ Y_i(1,0), & \text{if } Z_i = 1, W_i^{obs} = 0 \\ Y_i(1,1), & \text{if } Z_i = 1, W_i^{obs} = 1 \end{cases} \tag{1}$$

We designate the observed values Y_i^{obs} and W_i^{obs} to help distinguish them from their potential outcomes.

Several standard assumptions are required for valid inference. First, a user can only receive one version of the treatment and a user's treatment assignment does not interfere with other users' outcomes. This pair of assumptions is commonly known as the Stable Unit Treatment Value Assumption (SUTVA), a term coined in Rubin (1978). In our setting, this assumption is almost certainly satisfied due to Facebook's tracking abilities, which prevents those in the control condition from inadvertently being shown an ad. The second part of the SUTVA assumption could possibly be violated if users in the test group share ads with users in the control group, but we believe this happens rarely if at all.[13]

The second assumption is that assignment to treatment is unconfounded, or random. This requires that the distribution of Z_i is independent of all potential outcomes $Y_i(Z_i, W_i(Z_i))$ and both potential treatments $W_i(Z_i)$. Note that although assignment through Z_i is random, this does not imply treatment received W_i is random due to the one-sided non-compliance. The unconfoundedness assumption cannot formally be tested because we do not observe all potential outcomes and treatments, although we believe that Facebook implements its randomization procedure correctly. We will, however, perform a randomization check in each study to verify there are no differences in users' observed characteristics across test and control groups.

In principle, we could focus on the relationship between the random assignment Z_i and outcome Y_i, ignoring any information contained in W_i. Such an intent-to-treat (ITT) analysis only requires the two assumptions above. However, our interest is in estimating the causal effects of the receipt of treatment—what happens to users who are actually exposed to ads. For this we require an exclusion restriction:

$$Y_i(0, w) = Y_i(1, w), \text{ for all } w$$

such that assignment affects a user's outcome only through its receipt of the treatment. This assumption should be valid because users are unaware of their assignment to test or control, and so only exposure should affect outcomes. This permits Z_i to serve as an instrumental variable (IV) to recover the average treatment effect on the treated (ATT), which is our primary causal effect of interest.

[13] In theory, test users could show ads to control users, even though neither user knows their own assignment status. If this occurred, the treatment effect estimates would be conservative because this would only matter if a test user showed an ad to a control user who ended up converting.

3.2 Causal Effects in the RCT

Given the assumptions, the ITT effect of assignment on outcomes compares across random assignment status, irrespective of a user's treatment:

$$\text{ITT}_Y = \mathbb{E}\left[Y(1, W(1)) - Y(0, W(0))\right] \tag{2}$$

The sample ITT effect can be estimated using

$$\widehat{\text{ITT}}_Y = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1, W_i^{obs}) - Y_i(0, W_i^{obs})\right) \ , \tag{3}$$

which can written compactly as $\widehat{\text{ITT}}_Y = \overline{Y}_1^{obs} - \overline{Y}_0^{obs}$.

The ITT effect sidesteps issues with endogenous exposure and an advertiser might reasonably focus on this effect. However, oftentimes an advertiser wants to understand the effect of its ads on users who actually observed them because the firm might run another campaign that at Facebook to target this subpopulation or the firm might try to reach this group on other channels. For most of our analysis, we focus on the ATT, the causal effect for the subpopulation of exposed users.

Imbens and Angrist (1994) show this can be expressed in an IV framework as the ITT effect on the outcome divided by the ITT effect on the treatment receipt:

$$\tau_{att} = \frac{\text{ITT}_Y}{\text{ITT}_W} = \frac{\mathbb{E}[Y(1, W(1))] - \mathbb{E}[Y(0, W(0))]}{\mathbb{E}[W(1)] - \mathbb{E}[W(0)]} \tag{4}$$

With full compliance in the control, such that $W_i(0) = 0$ for all users, and complete randomization of Z_i, the denominator simplifies to $\text{ITT}_W = \mathbb{E}[W(1)]$.

Another way to think about this causal effect is to decompose the ITT effect for the entire sample into the weighted average of ITT effects for *compliers* and *noncompliers*. Compliers are users where $W_i(1) = 1$ is observed such that they comply with their assignment, and noncompliers are users with $W_i(1) = 0$ who do not comply with the assignment of $Z_i = 1$. The average ITT effect can be expressed as

$$\text{ITT}_Y = \text{ITT}_{Y,co} \cdot \pi_{co} + \text{ITT}_{Y,nc} \cdot (1 - \pi_{co}), \tag{5}$$

where $\pi_{co} = \mathbb{E}[W(1)]$ is the share of compliers. The exclusion restriction requires that $Y_i(0, 0) = Y_i(1, 0)$, which implies $\text{ITT}_{Y,nc} = \mathbb{E}\left[Y(1, 0) - Y(0, 0)\right] = 0$. Thus, $\text{ITT}_{Y,co}$ can be expressed as the overall ITT effect divided by the share of compliers

$$\text{ITT}_{Y,co} = \frac{\text{ITT}_Y}{\pi_{co}} \equiv \tau_{att} \ . \tag{6}$$

π_{co} can be estimated using the share of treated users in the test group. In a sense, scaling ITT_Y by the inverse of π_{co} "undilutes" the ITT effect according to the share of users who actually received

treatment in the test group (the compliers). These are the users who were induced to take up the treatment through assignment to the test condition.[14]

3.3 Lift

To help summarize outcomes across advertising studies, we report some results in terms of "lift", which is what Facebook uses internally.[15] Lift simply expresses the incremental conversion rate as a percentage effect:

$$\text{Lift} = \frac{\text{Actual conversion rate} - \text{Counterfactual conversion rate}}{\text{Counterfactual conversion rate}} \tag{7}$$

The lift for the ITT is

$$\text{Lift}_{ITT} = \frac{\overline{Y}_1^{obs} - \overline{Y}_0^{obs}}{\overline{Y}_0^{obs}} = \frac{\widehat{\text{ITT}}_Y}{\overline{Y}_0^{obs}} \tag{8}$$

Reporting the lift facilitates the comparison of advertising effects across studies because it normalizes the results using the baseline conversion rate in a study, which can vary significantly depending on the study's characteristics (e.g., the advertiser's identify, the outcome of interest, the period of study). The corresponding lift for the ATT is:

$$\text{Lift}_{ATT} = \frac{\mathbb{E}[W_i \cdot Y_i(1, W_i(1))] - (\mathbb{E}[W_i \cdot Y_i(1, W_i(1))] - \tau_{att})}{\mathbb{E}[W_i \cdot Y_i(1, W_i(1))] - \tau_{att}} = \frac{\tau_{att}}{\mathbb{E}[W_i \cdot Y_i(1, W_i(1))] - \tau_{att}} \tag{9}$$

One complication from using lift is that statistical inference must address the dependence between the terms in the ratio. Using the experimental data, we can implement a nonparametric bootstrap to estimate the sampling distribution of lift for both the ITT and ATT. However, for some of the observational models described in the next section, such an approach is either infeasible computationally or impossible due to nature of the method. In particular, Abadie and Imbens (2008) show the bootstrap is invalid as a inference technique for matching estimators. For this reason we conduct inference using the ITT and ATT effects, for which valid standard errors can be computed, and present lift estimates and standard errors where possible to help readers compare across studies. More details can be found in the Appendix.

4 Observational Models

This section is motivated around the following thought experiment. Rather than conducting a RCT, an advertiser followed customary practice by choosing a target sample and made all of these users

[14] Imbens and Angrist (1994) refer to this quantity as the local average treatment effect (LATE). If there are no "always-takers" and no "defiers" in the sample, which is true in our experimental design, the LATE is equal to the ATT.

[15] DellaVigna and Gentzkow (2010) advocate the reporting outcomes in terms of the persuasion rate.

eligible to see the ad. To estimate the treatment effect, the advertiser made use of the fact that some of the targeted sample had in fact not been exposed to the ad campaign. This is equivalent to creating a test sample without a control group held out. Below we describe the set of observational methods we apply to the test sample for comparison with the RCT results. All of these methods rely on some form of exogeneity, conditional on observables, that permits a user's treatment status to be considered as good as random.

To mimic this observational setting with the RCT data, we ignore the control group and focus exclusively on the test group. It is helpful to abuse notation slightly by defining:

$$Y_i(W_i) \equiv Y_i(Z_i = 1, W_i) \ . \tag{10}$$

For each user, we observe the triple (Y_i^{obs}, W_i, X_i), where the realized outcome is:

$$Y_i^{obs} \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0 \\ Y_i(1) & \text{if } W_i = 1 \end{cases} \tag{11}$$

If treatment status W_i were in fact random and independent of X_i, we could compare the conversion rates of exposed to unexposed users (Abraham 2008). The ATT effect would be:

$$\tau_{att}^{eu} = \mathbb{E}[Y_i(W_i = 1) - Y_i(W_i = 0)|X_i] = \mathbb{E}[Y_i(W_i = 1)] - \mathbb{E}[Y_i(W_i = 0)] \tag{12}$$

Replacing the expectations with their sample counterparts to estimate $\hat{\tau}_{att}^{eu}$ is straightforward. In reality, of course, W_i is unlikely to be independent of X_i, especially in the world of online advertising. The estimator $\hat{\tau}_{att}^{eu}$ will contain selection bias due to the dependence between user characteristics, treatment status, and outcomes. We report the lift based on $\hat{\tau}_{att}^{eu}$ in our results as a naive baseline.

Observational methods specifically attempt to correct for this selection bias. To accomplish this goal, all of the observational methods rely on two assumptions: unconfoundedness and overlap. The unconfoundedness assumption is

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \ \mid \ X_i, \tag{13}$$

which states that treatment status is independent of potential outcomes conditional on X_i. The second assumption concerns the joint distribution of treatment and covariates, and is frequently written as

$$0 < \Pr(W_i = 1|X_i) < 1$$

This assumption implies that each user in the population has some probability of being treated and some probability of being untreated. The probability of treatment as a function of x is known as the propensity score:

$$e(x) = \Pr(W_i = 1|X_i = x) = \mathbb{E}[W_i|X_i = x]. \tag{14}$$

Rosenbaum and Rubin (1983) show that if assignment to treatment is unconfounded conditional on X_i, then assignment is also unconfounded given the propensity score:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \;\mid\; e(X_i) \tag{15}$$

This assumption can be interpreted as, for two people with the same (or very close) propensity scores, exposure status is as good as random and independent of the potential outcomes. As in section 3.1, the unconfoundedness assumption in equation (13) is untestable. Overlap can be assessed before and after adjustments are made to each group, and which we will do in a future version of this paper.

In the following sections, we consider several classes of methods and some that combine elements of each. Recall that we already mentioned one naive benchmark that compares outcomes across exposed and unexposed users in the test group (equation 12). We adopt a second naive benchmark that is widely used in the industry, which is to adjust the exposed and unexposed groups based only on differences in the joint distribution of age and gender. We implement this procedure through Exact Matching (EM) on both variables, in effect matching all exposed users of a particular age-gender pair to all unexposed users with the same age and gender.

4.1 Regression

The fundamental problem is that we do not observe all potential outcomes for any given user. In some form or another, all methods impute the missing potential outcomes by building a model and using it to predict what would have happened to a user if the user had received a different treatment. Methods based on regression start with the following conditional expectation:

$$\mu_w(x) = \mathbb{E}[Y(w)|X = x] \tag{16}$$

Given unconfoundedness, this can be rewritten as

$$\mu_w(x) = \mathbb{E}[Y(w)|W = w, X = x] = \mathbb{E}[Y^{obs}|W = w, X = x] \tag{17}$$

which implies we can estimate $\mu_w(x)$ from the observed outcomes. Given consistent estimators $\hat{\mu}_w(x)$, the sample average treatment effect (ATE) using regression adjustment (RA) is:

$$\tau_{ate}^{ra} = \frac{1}{N}\sum_{i=1}^{N}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)] \tag{18}$$

The importance of the overlap assumption is evident because, for a given $X = x$, one must be able to estimate the conditional expectations at both values of the treatment. If the goal is to

estimate the average treatment effect of the treated, one must only estimate $\mu_0(x)$ to predict the counterfactual outcomes for the treated users:

$$\tau_{att}^{ra} = \frac{1}{N_T} \sum_{i=1}^{N} W_i [Y_i^{obs} - \hat{\mu}_0(X_i)] \tag{19}$$

where $N_T = \sum_i W_i$ is the number of treated units.

Any consistent estimator of $\mu_w(x)$ can be used, such as linear regression with covariates:

$$Y_i^{obs} = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i \tag{20}$$

Lewis, Rao, and Reiley (2011) use a logit regression after obtaining a control sample of unexposed users. In our implementation, we use separate regression functions for each treatment level

$$\mu_w(x) = \alpha_w + \beta_w' x \tag{21}$$

and include various interactions and higher-order terms as covariates.

One problem with these regression estimators is their sensitivity to differences in the covariate distributions for control and test groups. If these distributions differ, these estimators will rely heavily on extrapolation. Researchers have gone beyond simple parametric regression models, with methods that rely on local smoothing, such as kernel methods or locally linear regression (Heckman, Ichimura, and Todd 1997, Heckman, Ichimura, and Todd 1998, Heckman, Ichimura, Smith, and Todd 1998), and flexible global approximations, such as series estimators (Hahn 1998, Imbens, Newey, and Ridder 2005). However, applying these methods relies on the optimal choice of a smoothing parameter, or bandwidth, to determine the extent of local extrapolation. With the exception of Imbens, Newey, and Ridder (2005), most work relies on ad hoc methods for this choice.

4.2 Matching on the Propensity Score

Matching methods try to pair users in one treatment group with "similar" users in the opposite treatment group. The treatment effect is estimated as the difference between an observation and its paired observation(s). The result is a new matched sample of observations that hopefully have more similar covariate distributions compared to the unmatched sample. Averaging over the relevant set of observations—the entire sample or only treated users—yields the relevant sample treatment effects. Thus, matching methods address a shortcoming of regression methods that may require extrapolation from the covariate distribution in one treatment group to the estimate the outcomes in the other treatment group. Fundamentally, though, the two approaches are similar: both impute the missing potential outcomes from one group using outcomes in the other. In regression, the

missing potential outcomes come from the estimated regression function. Matching methods also impute missing outcomes but use the nearest neighbors from the opposite treatment group, defined by some similarity metric. This approach can be viewed as being similar to a nonparametric kernel with the number of neighbors serving as the fixed bandwidth.

Matching entails a variety of choices on the part of the econometrician: which similarity metric to use, whether to match with or without replacement, and how many matched units to use. If X_i is entirely discrete, one could in principle match users exactly. Given that the dimensionality of X_i is usually large and contains a mixture of discrete and continuous variables, exact matching on all covariates is typically infeasible even with large samples. Various metrics have been proposed, with the most common being the Mahalanobis distance and the propensity score (Rosenbaum and Rubin 1985, Deheji and Wahba 2002, Caliendo and Kopeinig 2008, Rubin and Thomas 2000). In both cases, the goal is to summarize the information contained in the full set of covariates to facilitate the construction of matched test and control groups with similar covariate distributions.

We match users based on their estimated propensity scores. Define $\ell(x) = \ln(e(x)/(1-e(x)))$ as the log-odds ratio of the propensity score. We use $\ell(x)$ rather than $e(x)$ because the former linearizes values on the unit interval. The relevant distance metric we consider is:

$$d_\ell(x, x') = \big(\ell(x) - \ell(x')\big)^2 \ ,$$

where we replace $e(x)$ with $\hat{e}(x)$ for estimation. We estimate the propensity scores using a logistic regression:

$$e(x) = \frac{\exp(x')}{1+\exp(x')},$$

For each user in the exposed group, we find M users in the unexposed group. We implement this matching procedure with replacement because it can reduce the bias of the estimator and is less computationally burdensome to implement. In the studies we consider at Facebook, N_T can be quite large and so the computational advantages of matching with replacement can be important. Two downsides of using replacement are that the sampling variance might be larger and estimating the sampling variance is more difficult (Abadie and Imbens 2015).

To implement the estimator, let $m_i^{c,k} \in \mathbb{I}_c$ be the index of the control unit that is the k^{th} closest to exposed user i based on the distance metric $d_\ell(x_i, x_k)$. The set $\mathcal{M}_i^c = \{m_i^{c,1}, m_i^{c,2}, \ldots, m_i^{c,M}\}$ contains the M closest observations for user i. For exposed user i, we observe $Y_i^{obs} = Y_i(1)$, and so we require an estimate of the potential outcome $Y_i(0)$. The counterfactual estimate of this outcome is

$$\widehat{Y_i(0)} = \frac{1}{M} \sum_{j \in \mathcal{M}_i^c} Y_j^{obs} \ . \tag{22}$$

The matching estimator for the average treatment effect on the treated using the estimated propensity scores is

$$\hat{\tau}_{att}^{psm} = \frac{1}{N} \sum_{i \in \mathbb{I}_c} \left(Y_i(1) - \widehat{Y_i(0)} \right) \ . \tag{23}$$

4.3 Blocking and Regression

Matching and regression can be combined in various ways. One approach, advocated by Imbens and Wooldridge (2009) and others, is block the data on the estimated propensity score (also known as subclassification or stratification) and using regression within blocks to estimate the causal effect. The idea is that the covariate distribution within a block should be relatively balanced, such that a regression should not have to overly rely on extrapolation across the test and control groups. After estimating the propensity score, the sample is divided into blocks (or strata) such that within each block the estimated propensity scores are approximately constant. The causal effect could be estimated within each block as if assignment was actually random within the block. We can go one step further by applying regression to help correct for any remaining imbalances within a block.

This estimator begins by partitioning the range of the propensity score into M intervals of $[b_{j-1}, b_j)$, for $j = 1, \ldots, J$, where $b_0 = 0$ and $b_J = 1$. One way to implement this is to divide the unit interval into equispaced blocks with boundary values at m/M for $m = 1, \ldots, M-1$. Let $B_i(m)$ be a binary indicator that user i is contained in block m, defined as

$$B_i(m) = 1 \cdot \left\{ \frac{m-1}{M} < e(X_i) \leq \frac{m}{M} \right\} \tag{24}$$

for $m = 1, \ldots, M$. Each block contains N_{wm} observations with treatment $w \in \{0, 1\}$, $N_{wm} = \sum_i 1\{W_i = w\} B_i(m)$. For each subgroup, we estimate the average treatment effect as if random assignment held within that subgroup,

$$\hat{\tau}_m = \frac{1}{N_{1m}} \sum_{i=1}^{N} B_i(m) W_i Y_i - \frac{1}{N_{0m}} \sum_{i=1}^{N} B_i(m)(1 - W_i) Y_i \ . \tag{25}$$

The overall average treatment effect can be estimated by calculating the weighted average of the block-specific treatment effects according to the number of users within each block relative to the full sample. Similarly, the average treatment effect of the treated is the average using weights equal to the proportion of treated users in each block,

$$\hat{\tau}_{block} = \sum_{m=1}^{M} \hat{\tau}_m \frac{N_{1m}}{N_T} \tag{26}$$

Blocking can be combined with regression for added flexibility. Within a block, we can re-express the treatment effect using the least squares estimator of τ_m in the regression

$$Y_i = \alpha_m + \tau_m \cdot W_i + \beta_m' X_i + \varepsilon_i. \tag{27}$$

using only individuals in block m. As in equation (26), this produces a set of M estimates which can be averaged appropriately to calculate the average treatment on the treated, and the estimates are equivalent if covariates are omitted from 27. The benefit of incorporating regression is that blocking on the propensity score should have already helped to balance approximately the covariates within each block. This implies the block-specific regressions rely less on extrapolation to estimate the treatment effect but can still adjust for any remaining covariate imbalance within the block.

One question is how many blocks to use. Assuming unconfoundedness, all the bias is due to the propensity score. With a single normally distributed covariate, five blocks removes most of the bias (Cochran 1968). More recent methods, developed in Imbens and Rubin (2014), follow a data-driven procedure to determine the optimal number of blocks and their boundaries, and where the number of blocks increases with the sample size.

4.4 Inverse Probability-Weighed Regression Adjustment

Another approach combines regression and the propensity score in a different manner. Rather than matching on the propensity score, we use it to form weights to help control for correlation between treatment status and the covariates. This method belongs to a class of procedures that have the "doubly robust" property (Robins and Ritov 1997). This means the estimator for the ATT (or ATE) is consistent even if one of the underlying models—either the propensity model or the outcome model—turns out to be misspecified.

Suppose the propensity score is modeled using $e(x) = \rho(x;\gamma)$, with $\hat{\gamma}$ estimated by maximum likelihood to obtain the estimated propensity scores $\hat{e}(X_i) = \rho(X_i, \hat{\gamma})$. Next we weight the objective function of the outcome model, in this case a linear regression, by the inverse probability of treatment or non-treatment.

$$\min_{\{\alpha_w,\beta_w\}} \sum_{i=1}^{N} (1 - W_i) \frac{(Y_i^{obs} - \alpha_0 + \beta_0' X_i)^2}{1 - \rho(X_i, \hat{\gamma})} + W_i \frac{(Y_i^{obs} - \alpha_1 + \beta_1' X_i)^2}{\rho(X_i, \hat{\gamma})}$$

This method is known as inverse-probability-weighed regression adjustment (IPWRA) model. In practice, we use a logit model for both the outcomes and propensity models, and the estimation procedure combines the two steps in a GMM specification to help calculate standard errors (Wooldridge 2007).

5 Data

The 12 advertising studies analyzed in this paper were chosen by two of the authors (Gordon and Zettelmeyer) for their suitability for comparing several common ad effectiveness methodologies and

Table 1: Summary statistics for all studies

Study	Vertical	Observations	Test	Control	Impressions	Clicks	Conversions	Outcomes*
1	Retail	2,427,494	50.0%	50.0%	39,167,679	45,401	8,767	C, R
2	Finan. serv.	86,183,523	85.0%	15.0%	577,005,340	247,122	95,305	C, P
3	E-commerce	4,672,112	50.0%	50.1%	7,655,089	48,005	61,273	C
4	Retail	25,553,093	70.0%	30.0%	14,261,207	474,341	4,935	C
5	E-commerce	18,486,000	50.0%	50.0%	7,334,636	89,649	226,817	C, R, P
6	Telecom	141,254,650	75.0%	25.0%	590,377,329	5,914,424	867,033	P
7	Retail	67,398,350	17.0%	83.0%	61,248,021	139,471	127,976	C
8	E-commerce	8,333,319	50.0%	50.1%	2,250,984	204,688	4,102	C, R
9	E-commerce	71,068,955	75.0%	25.0%	35,197,874	222,050	113,531	C
10	Tech	1,955,375	60.0%	40.0%	2,943,890	22,390	7,625	C, R
11	E-commerce	13,339,044	50.0%	50.0%	11,633,187	106,534	225,241	C
12	Finan. serv.	16,578,673	85.0%	15.0%	23,105,265	173,988	6,309	C

* C = checkout, R = registration, P = page view

for exploring the problems and complications of each. All 12 studies were randomized controlled trials held in the US. The studies are not representative of all Facebook advertising, nor are they intended to be representative. Nonetheless, they cover a varied set of verticals (retail, financial services, e-commerce, telecom, and tech). Each study was conducted recently (January 2015 or later) on a large audience (at least 1 million users) and with conversion tracking implemented by the advertiser. All studies restrict attention to users aged 18 and older.

5.1 Descriptive Statistics and Covariates

Table 1 provides summary statistics for each study. The studies range in size, with the smallest containing around two million users and the largest about 140 million. There is a mix of test/control splits. The studies also differed by the conversion outcome that the advertiser measured; some advertisers tracked multiple outcomes of interest. In all studies but one, the advertiser placed a conversion pixel on the checkout confirmation page, therefore gaining the ability to measure whether a Facebook user purchased from the advertiser. In four studies the advertiser placed a conversion pixel to measure whether a consumer registered with the advertiser. In three studies the advertiser placed a conversion pixel on a (landing) page of interest to the advertiser (termed a "key page view").

Table 2 provides brief information on the variables we observe. Eleven of the studies contain each of the variables in this table, but for one study we only observe variables in group 1. For most of the observational models, we implement a sequence of specifications corresponding to the following grouping of covariates:

1. The first specification includes group 1 variables from Table 2, which are common Facebook

Table 2: Description of Variables

Group	Variable	Description	Source
1	age	Age of user	FB
1	gender	1 = female, 0 = male	FB
1	married	1 = user is married	FB
1	single	1 = user is single	FB
1	inrelationship	1 = user is in a relationship	FB
1	engaged	1 = user is engaged	FB
1	FB age	Days since user joined FB	FB
1	friends	# of friends	FB
1	num_initiated	# of friend requests sent	FB
1	L7	# of last 7 days accessed FB	FB
1	L28	# of last 28 days accessed FB	FB
1	web_L7	# of last 7 days accessed FB by desktop	FB
1	web_L28	# of last 28 days accessed FB by desktop	FB
1	mobile_L7	# of last 7 days accessed FB by mobile	FB
1	mobile_L28	# of last 28 days accessed FB by mobile	FB
1	mobile_phone_OS	operating system of primary phone	FB
1	region	region of user's residence	FB
2	population	population in zip code	ACS
2	housingunits	# of housing units	ACS
2	pctblack	% black residences	ACS
2	pctasian	% asian residences	ACS
2	pctwhite	% white residences	ACS
2	pcthisp	% hispanic residences	ACS
2	pctunder18	% residents under age 18	ACS
2	pctmarriedhh	% married households	ACS
2	yearbuilt	average year residences built	ACS
2	pcths	% residents with at most high school degree	ACS
2	pctcol	% residents with at most college degree	ACS
2	pctgrad	% residents with graduate degree	ACS
2	pctbusfinance	% working in business/finance	ACS
2	pctstem	% workign in STEM	ACS
2	pctprofessional	% working in professional jobs	ACS
2	pcthealth	% working in health industry	ACS
2	pctprotective	% working in protectice services	ACS
2	pctfood	% working in food industry	ACS
2	pctmaintenance	% working in maintenance	ACS
2	pcthousework	% working in home services	ACS
2	pctsales	% working in sales	ACS
2	pctadmin	% working in adminstration	ACS
2	pctfarmfish	% working at farms or fisheries	ACS
2	pctconstruction	% working in construction	ACS
2	pctrepair	% working in repair industry	ACS
2	pctproduction	% working in production indutry	ACS
2	pcttransportation	% working in transportation industry	ACS
2	income	average household income	ACS
2	medhhsize	median household size	ACS
2	medhvalue	median household value	ACS
2	vehperh	average vehicles per household	ACS
2	pcthowned	% households who own a home	ACS
2	pctvacant	% vacant residences	ACS
2	pctunemployed	% unemployment	ACS
2	pctbadenglish	% residents with "bad" english	ACS
2	pctpoverty	% residents living below poverty line	ACS
3	match_score	Composite variable of FB data	FB

First seven rows are self-reported by the users. ACS data is from 2010.

variables such as age, gender, how long users have been on Facebook, how many Facebook friends the have, their reported relationship status, their phone OS, and other user characteristics.

2. In addition to the variables in 1, this specification uses Facebook's estimate of the user's zip code of residence to associate with each user nearly 40 variables drawn from the most recent Census and American Communities Surveys (ACS).

3. In addition to the variables in 2, this specification adds a composite metric of Facebook data that summarizes thousands of behavioral variables. This is a machine-learning based metric used by Facebook to construct target audiences that are similar to consumers that an advertiser has identified as desirable.[16] Using this metric bases the estimation of our propensity score on a non-linear machine-learning model with thousands of features.[17]

5.2 Randomization Checks

An important step is to check whether the randomization was implemented correctly. Table 3 provides evidence that several variables have comparable means across the test and control groups for one typical study (study 4).

To summarize this information across studies, we compared means across test and control for each study and variable, resulting in 624 p-values. Of these, 10.4% are below 0.10, 5.1% are below 0.05, and 0.9% are below 0.01. Under the null hypothesis that the means are equal, the resulting p-values from the hypothesis tests should be uniformly distributed on the unit interval. Figure 4 shows this is indeed the case. We have also looked unsuccessfully for any evidence that particular variables might be more likely to exhibit imbalance. Thus, based on this collection of results, we fail to find any evidence that the randomization was implemented improperly.

6 Results for Study 4

Before presenting the finding across all the studies, in this section we walk through the results in detail for a typical advertising study (we refer to it as "study 4"). To preserve confidentiality, all of

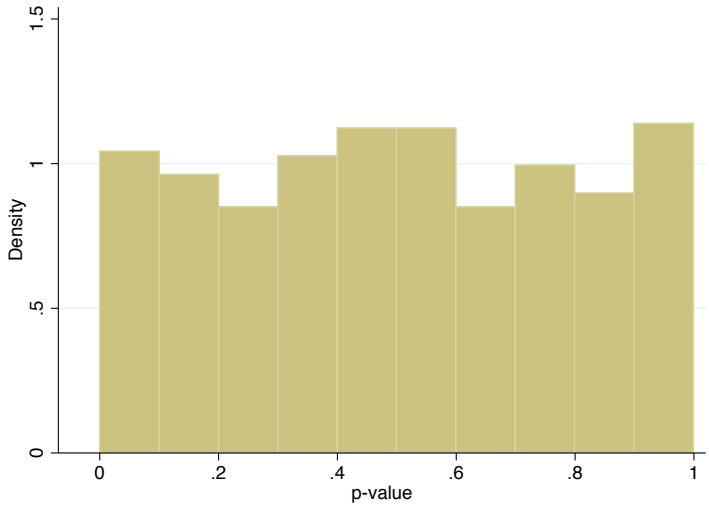
[16] See `https://www.facebook.com/business/help/164749007013531` for an explanation.

[17] Please note that, while this specification contains a many user-level variables, in this version of the paper we have no data at the user level that varies over time within the duration of each study. For example, while we know whether a user used Facebook during the week prior to the beginning of the study, we don't observe on any given day of the study whether the user used Facebook on the previous day or whether the user engaged in any shopping activity. It is possible that using such time-varying user-level information could improve our ability to match. We hope to explore this in a future version of the paper.

Table 3: Randomization check for study 4

Variable	Control group	Test group	p-value
Average user age	31.7	31.7	0.33
% of users who are male	17.2%	17.2%	0.705
Length of time using FB (days)	2,288	2,287	0.24
% of users with status "married"	19.6	19.6	0.508
% of users status "engaged"	13.8	13.8	0.0892
% of users status "single"	14.0	14.0	0.888
# of FB friends	485.7	485.7	0.985
# of FB uses in last 7 days	6.377	6.376	0.14
# of FB uses in last 28 days	25.5	25.5	0.172

Figure 4: Distribution of p-values across all studies



the conversion rates in this section and in section 7 have been scaled by a random constant, such that relative comparisons across studies are valid but their absolute levels have been masked.

Study 4 was performed for the advertising campaign of an omni-channel retailer. The campaign took place over two weeks in the first half of 2015 and comprised a total of 25.5 million users. Ads were shown on mobile and desktop Facebook news feeds in the US. For this study the conversion pixel was embedded on the checkout confirmation page. The outcome measured in this study is whether a user purchased online during the study and up to several weeks after the study ended.[18] Users were randomly split into test and control groups in proportions of 70%, and 30%, respectively.

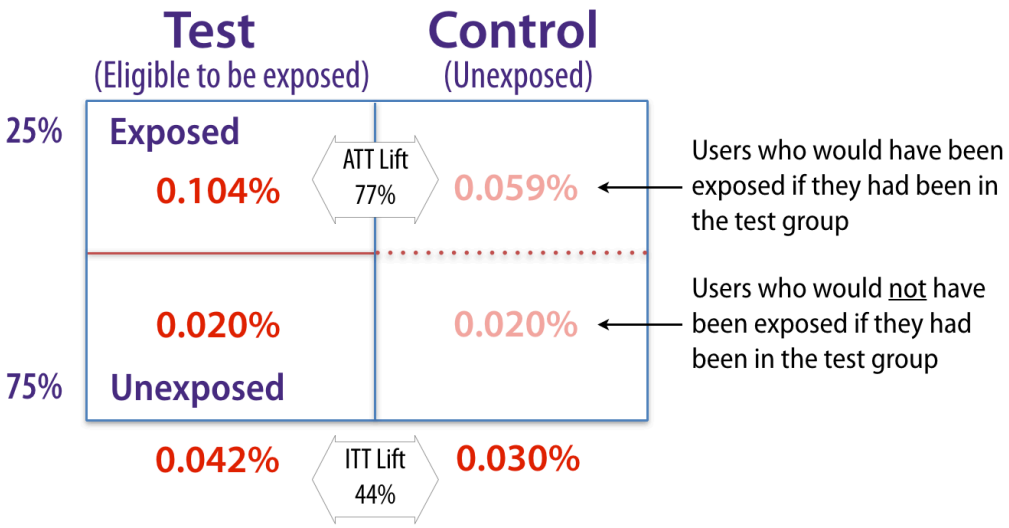
[18]Even if some users convert as a result of seeing the ads further in the future, this still implies the experiment will produce conservative estimates of advertising's effects.

6.1 RCT

Figure 5 summarizes the results from the RCT. The conversion rates in the control and test group were 0.030% and 0.042%, respectively, implying an ITT lift of 44%. The estimated ATT was 0.045%. Based on the conversion rate of 0.104% for treated users in the test group, this implies the ATT lift was 77% (=0.045%/0.059%). The 95% confidence interval for this lift is [37%, 117%].[19] Note that we do not actually know which users in the control group would have been exposed had they been assigned to the test group—this represents one of the basic challenges for estimating the ATT.

We will interpret the 77% lift measured by the RCT as our gold standard measure of the truth. In the following subsections we will calculate alternative measures of advertising effectiveness to see how close they come to this 77% benchmark. These comparisons reveal how close to (or far from) knowing the truth an advertiser who was unable to (or chose not to) evaluate their campaign with an RCT, would be.

Figure 5: Results from RCT



6.2 Observational Models

In the current version of the paper, we present estimates using Exact Matching (EM) based only on age and gender (as a naive benchmark widely used in industry), propensity score matching (PSM), and inverse probability weighted regression adjustment (IPWRA). In the next version, we will add

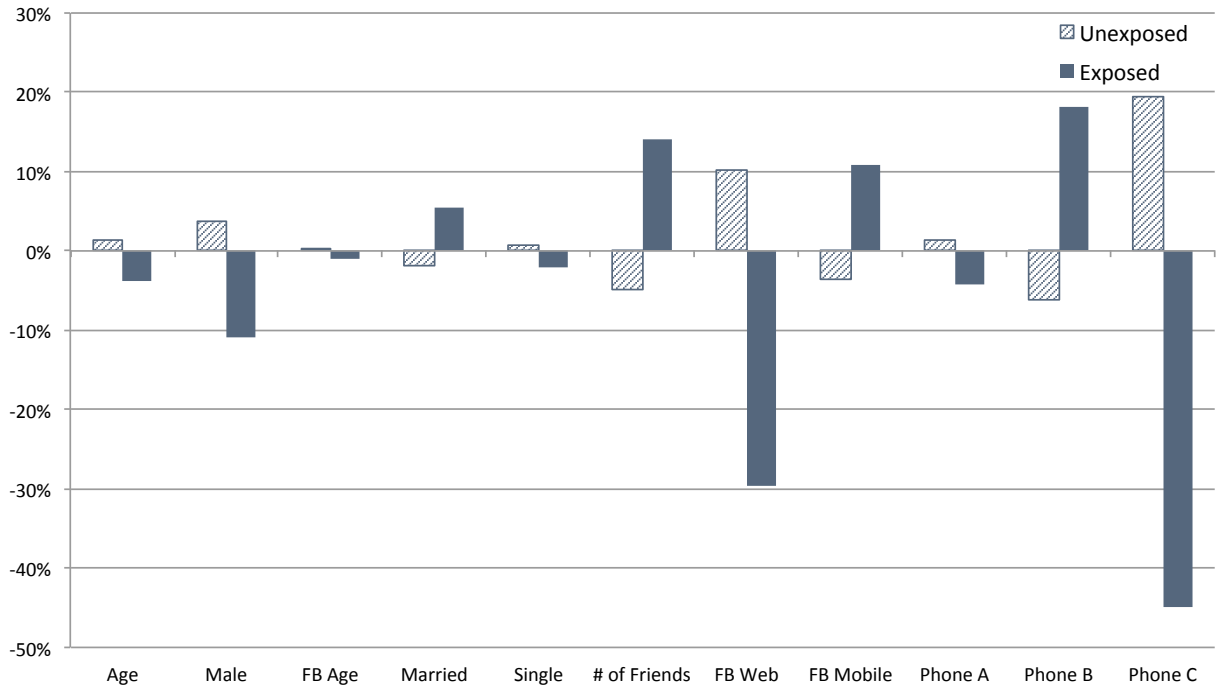
[19] See the technical appendix for details on how to compute the confidence interval for the lift.

simple regression adjustment (RA), Blocking/Regression, and provide additional robustness checks for each method.

Before delving into the observational models, it will be useful to characterize some of the selection bias present in this study. Figure 6 depicts the differences between the two groups in study 4. The percentage differences are relative to the average of the entire test group. For example, the second item in Figure 6 shows that exposed users are about 10% less likely to be male than the average for the test group as a whole, while unexposed users are several percentage points more likely to be male than the average for the whole group. The figure also shows that exposed users are more likely to be female, are slightly younger, are more likely to be married, have more Facebook friends, and tend to access Facebook more frequently from a mobile device than a desktop.

If we are willing to (incorrectly) assume that exposure is random, we could compare the exposed and unexposed groups, as in equation (12). The conversion rate among exposed users was 0.104% and the conversion rate among unexposed users was 0.020%, implying an ATT lift of 416%. This estimate is more that five times the true lift of 77%.

Figure 6: Comparison of exposed and unexposed users in the test group of study 4
(expressed as percentage differences relative to average of the entire test group)



EM. Within the test group of study 4, there were 113 unique combinations of age and gender

for which there was at least one exposed and at least one unexposed user. Seven age-gender combinations were dropped due to a lack of overlap, reducing the sample by only 15 users. The remaining unexposed users are matched to the exposed users by age-pair value, leading to some amount of re-weighting among the unexposed users. The (unweighted) exposed users converted at a rate of 0.104% and (weighted) unexposed users at 0.032%, for a lift of 221%. This estimate is roughly half the lift obtained from directly comparing exposed and unexposed users but still much greater than the RCT lift of 77%. This remaining bias is not surprising given the differences in user characteristics evident in Figure 6.

PSM. Table 4 presents a summary of the estimates of advertising effectiveness produced by the exact matching and propensity score matching approaches. As before, the main result of interest will be the lift. In the context of matching models, lift is calculated as the difference between the conversion rate for matches exposed users and matched unexposed users, expressed as a percentage of the conversion rate for matched unexposed users. Table 4 reports each of the components of this calculation, along with the 95% confidence interval for each estimate. The bottom row reports the AUCROC, a common measure of the accuracy of classification models (it applies only to the propensity score models).[20]

Note that the conversion rate for matched exposed users barely changes across the model specifications. This is because for the most part we are holding on to the entire set of exposed users and changing across specifications which unexposed users are chosen as the matches.[21] Consequently, the conversion rate of the matched unexposed users changes across specification. This is because different specifications choose different sets of matches from the unexposed group, When we go from exact matching (EM) to our most parsimonious propensity score matching model (PSM 1), the conversion rate for unexposed users increases from 0.032% to 0.042%, decreasing the implied advertising lift from 221% to 147%. PSM 2 performs similarly to PSM 1, with an implied lift of 154%.[22] Finally, adding the composite measure of Facebook variables in PSM 3 improves the fit of the propensity model (as measured by a higher AUCROC) and further increases the conversion rate for matched unexposed users to 0.051%. The result is that our best performing PSM model estimates an advertising lift of 102%.

[20] See `http://gim.unmc.edu/dxtests/roc3.htm` for a short and Fawcett (2006) for a detailed an explanation of AUCROC.

[21] Exposed users are dropped if there is no unexposed user that has a close enough propensity score match. In study 4, the different propensity score specifications we use do not produce very different sets of exposed users who can be matched. This need not be the case in all settings.

[22] As we add variables to the propensity score model, we must drop some observations in the sample with missing data. However, the decrease in sample size is fairly small and these dropped consumers do not significantly differ from the remaining sample.

Table 4: Exact Matching (EM) and Propensity Score Matching (PSM 1-3)

	EM		PSM 1		PSM 2		PSM 3	
	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
Conversion rates for matched unexposed users (%)								
	0.032	[0.029, 0.034]	**0.042**	[0.041, 0.043]	**0.041**	[0.040, 0.042]	**0.051**	[0.050, 0.052]
Conversion rates for matched exposed users (%)								
	0.104	[0.097, 0.109]	**0.104**	[0.097, 0.109]	**0.104**	[0.097, 0.109]	**0.104**	[0.097, 0.110]
Lift (%)	**221**	[192, 250]	**147**	[126, 168]	**154**	[132, 176]	**102**	[83, 121]
AUCROC	N/A		0.72		0.73		0.81	
Observ	7,674,114		7,673,968		7,608,447		7,432,271	

*Slight differences in the number of observations are due to variation in missing characteristics across users in the sample. Note that the confidence intervals for PSM 1-3 on the conversion rate for matched unexposed users and the lift are approximate (consult the appendix for more details).

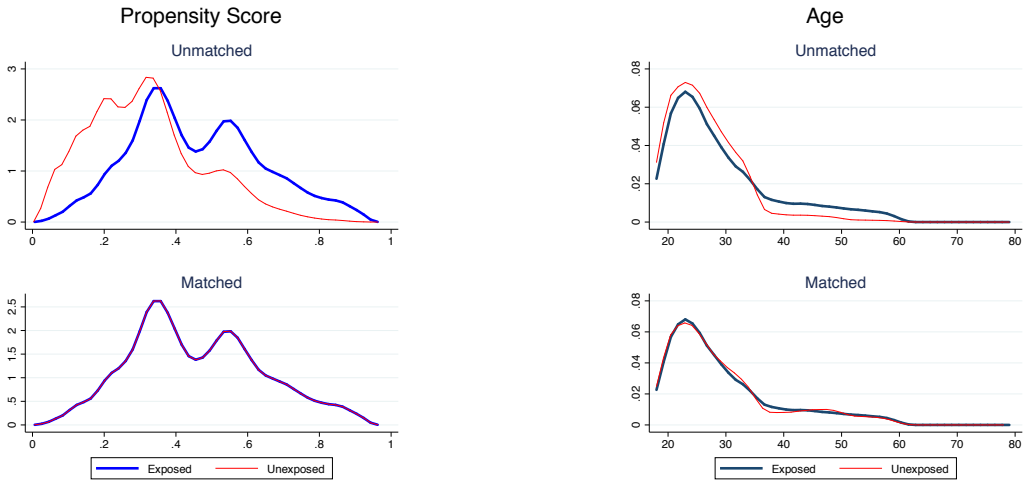
We assessed how well matching on propensity score balanced the exposed and unexposed groups. The upper part of panel (a) of Figure 7 shows the distributions of the propensity scores for all exposed and unexposed users prior to matching. The lower part of panel (a) shows the distributions of the matched sample. Prior to matching, the propensity score distribution for the exposed and unexposed users differ substantially. After matching, however, there is no visible difference in the distributions, implying that matching did a good job of balancing the two groups based on their likelihood of exposure.[23] Propensity score matching matches users based on a composition of their characteristics. One might wonder how well propensity-score matched samples are matched on individual characteristics. In panel (b) of Figure 7, we show the distribution of age for exposed and unexposed users in the unmatched samples (upper) and in the matched samples (lower). Even though we did not match directly on age, matching on the propensity score nevertheless balanced the age distribution between exposed and unexposed users.

IPWRA. We estimated three different regression adjustment models. Table 5 presents the results. The results are similar to those obtained using PSM: including additional variables reduces the estimated lift from 145% to 107%.

6.3 Summary of Results for Study 4

Figure 8 summarizes the results from all methods applied to Study 4. When we naively compared exposed to unexposed users, we estimated an ad lift of 416%. Adjusting these groups to achieve

[23]This comparison also helps us check that we have sufficient overlap in the propensities between the exposed and unexposed groups.



(a) Propensity Score (b) Age

Figure 7: Comparison of Unmatched and Matched Characteristic Distributions

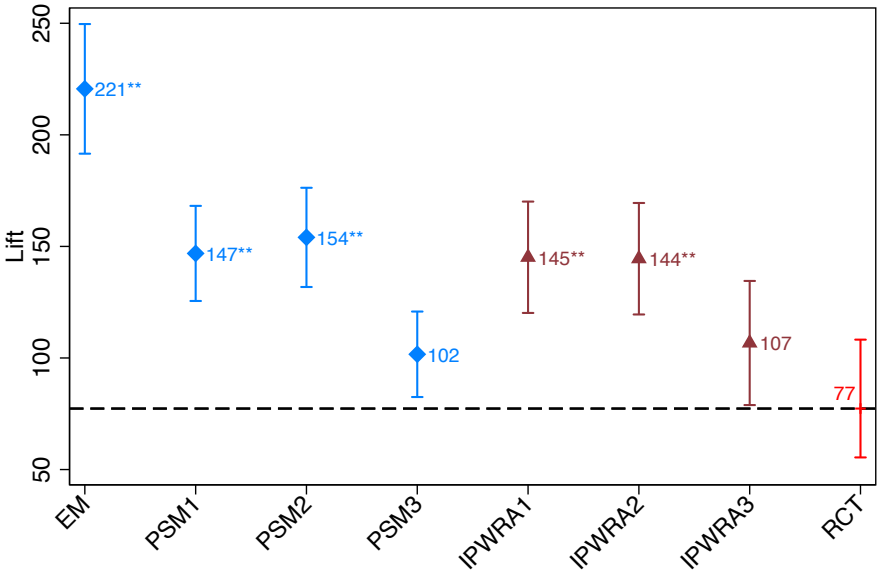
Table 5: Inverse-probability-weighted Regression Adjustment

	IPWRA 1		IPWRA 2		IPWRA 3	
	Est.	CI	Est.	CI	Est.	CI
Conversion rate for exposed users if unexposed as predicted by RA mdoel (%)						
	0.045	[0.037, 0.046]	**0.045**	[0.039, 0.046]	**0.049**	[0.044, 0.056]
Actual conversion rate of exposed users						
	0.104	[0.097, 0.109]	**0.102**	[0.096, 0.107]	**0.104**	[0.097, 0.110]
Lift%	**145**	[120, 171]	**144**	[120, 169]	**107**	[79, 135]

balance on age and gender alone through exact matching yields a lift of 221%. Matching the groups based on their propensity score, estimated with a rich set of explanatory variables, gave us a lift of 102%. Compared to the starting point, we have gotten much closer to the true RCT lift of 77%.

As the figure shows, propensity score matching and regression methods perform comparably well. Both methods tend to overstate lift, although including our complete set of predictor variables—especially the composite Facebook variable—produce lift estimates that are statistically indistinguishable from the RCT lift. However, if one ignores the uncertainty represented in confidence intervals and focuses on the point estimates alone, even a model with a rich set of predictors overestimates the lift by about 50%.

Figure 8: Summary of lift estimates and confidence intervals for Study 4



** (*) Statistically significantly different from RCT lift at 1% (5%) level.

7 Results From All 12 Studies

In section 6 we presented the results from applying a variety of observational approaches to estimate the lift of study 4 in comparison to the RCT estimate. In this section we summarize the findings of using the same approaches for all 12 studies.

7.1 RCT Results

Table 6 presents the results of the RCTs for all studies. A reasonable amount of variation exists across studies in the percentage of the test group who are exposed to ads and in the ATT lift. Of the 11 studies with a checkout conversion, three failed to produce statistically significant lifts.

The lifts for registration and page view outcomes are typically higher than for checkout outcomes. The reason is as follows: Since specific registration and landing pages are typically tied to ad campaigns, users who are not exposed to an ad are much less likely to reach that page than users who see the ad, simply because unexposed users may not know how to get to the page. For checkout outcomes, however, users in the control group lead to a checkout outcome simply by purchasing from the advertiser—it does not take special knowledge of a page to trigger a conversion pixel.[24]

[24]One might ask why lifts for registration and page view outcomes are not infinite since—as we have just claimed—users only reach those pages in response to an ad exposure. The reason is that registration and landing pages are often shared among several ad campaigns. Therefore, users who are in our control group might have been exposed to a different ad campaign which shared the same landing or registration page.

Table 6: Lift for all studies and measured outcomes

Study	Outcome	Pct Exposed	RCT ATT Lift	Confidence Interval	
1	Checkout	76%	33%	[19.5%	48.9%]
2	Checkout	46%	0.91%	[-4.3%	7.2%]
3	Checkout	63%	6.9%	[0.02%	14.3%]
4	Checkout	25%	77%	[55.4%	108.2%]
5	Checkout	29%	418%	[292.8%	633.5%]
7	Checkout	49%	3.5%	[0.6%	6.6%]
8	Checkout	26%	-3.6%	[-20.7%	19.3%]
9	Checkout	6%	2.5%	[0.2%	4.8%]
10	Checkout	65%	0.6%	[-13.8%	16.3%]
11	Checkout	40%	9.8%	[5.8%	13.8%]
12	Checkout	21%	76%	[56.1%	101.2%]
1	Registration	65%	789%	[696.0%	898.4%]
5	Registration	29%	900%	[810.0%	1001.9%]
8	Registration	29%	61%	[12.3%	166.1%]
10	Registration	58%	8.8%	[0.4%	18.2%]
2	Page View	76%	1617%	[1443.8%	1805.2%]
5	Page View	46%	601%	[538.6%	672.3%]
6	Page View	26%	14%	[12.9%	14.9%]

RCT Lift in red: statistically different from zero at 5% level. 95% confidence intervals obtained via bootstrap.

7.2 Observational Models Results

We summarize the results of the exact matching specification (EM), the three propensity score matching specifications (PSM 1-3), and the three regression adjustment specifications (IPWRA 1-3) using the same graphical format with which we summarized study 4 (see Figure 8). Figures 9 and 10 summarize results for the eleven studies for which there was a conversion pixel on the checkout confirmation page.

- In **study 1**, the exact matching specification (EM), the first two propensity score matching specifications (PSM 1 and 2), and the first two inverse-probability-weighed regression adjustment specifications (IPWRA 1 and 2) yield lift estimates between 85% and 117%, which are statistically higher than the RCT lift of 33%. Including the composite metric of Facebook data that summarizes thousands of behavioral variables (PSM 3 and IPWRA 3) lowers the lift estimate to 59%, which is not statistically different from the RCT lift. Hence, study 1 shows a similar pattern to the one we observed in study 4.

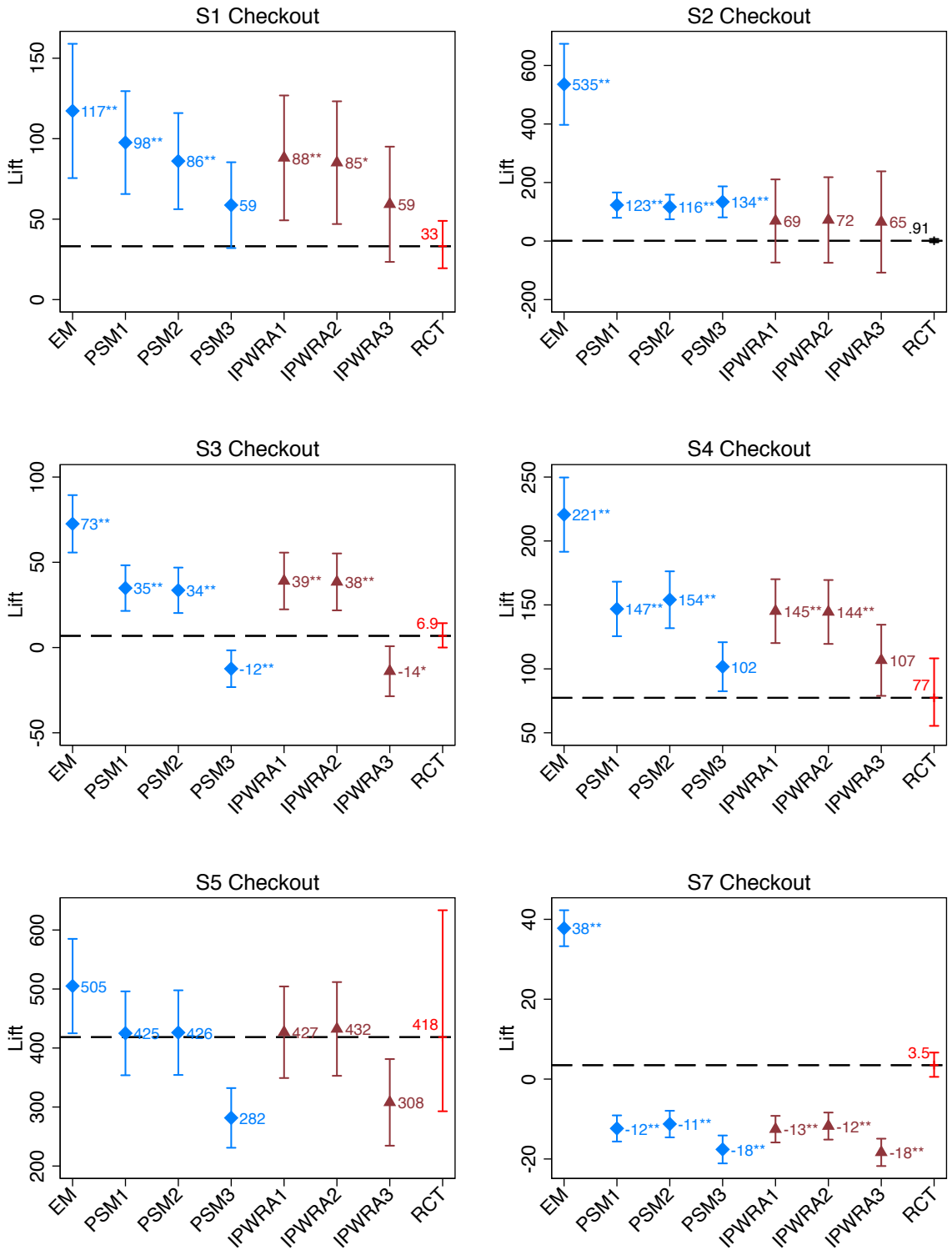
- The results for **study 2** look very different. The RCT shows no significant lift. Nonetheless, the EM and all PSM specifications yield lift estimates of 116 to 535%, all of which are statistically higher than the RCT estimate of 0.91%. The lift estimates of the IPWRA specifications are between 65 and 72%, however, they are also very imprecisely measured and

therefore statistically not different from the RCT estimate.

- **Study 3** follows yet another pattern. The RCT lift is 6.9%. EM, PSM 1, PSM 2, IPWRA 1, and IPWRA 2 all overestimate the lift (34-73%). PSM 3 and IPWRA 3, however, significantly underestimate the RCT lift (-12 to -14%).
- **Study 4** was already discussed in section **??**.
- In **study 5** all estimates are statistically indistinguishable from the RCT lift of 418%. The point estimates range from 515% for EM to 282% for PSM 3.
- **Study 6** did not feature a checkout conversion pixel.
- In **study 7** all estimates are different from the RCT lift of 3.5%. EM overestimates the lift with an estimate of 38%. All other methods underestimate the lift with estimates between -11 and -18%.
- Moving to Figure 10, **study 8** finds an RCT lift of -3.6% (not statistically different from 0). All methods overestimate the lift with estimates of 23 to 49%, except for IPWRA3 with a lift of 16%, which is not statistically different from the RCT lift.
- The RCT lift in **study 9** is 2.5%. All observational methods massively overestimate the lift; estimates range from 1413 to 3288%.
- **Study 10** estimates an RCT lift of 0.6% (not statistically different from 0). The point estimates of different methods range from -18 to 37%, however, only the EM lift estimate (37%) is statistically different from the RCT lift.
- **Study 11** estimates an RCT lift of 9.8%. EM massively overestimates the lift at 276%. PSM 1, PSM 2, IPWRA 1, and IPWRA 2 also overestimate the lift (22-25%), but to a much smaller degree. PSM 3 and IPWRA 3, however, estimate a lift of 9.4 and 3.5%, respectively. The latter estimates are not statistically different from the RCT lift. PSM 3 in study 11 is the only case in these 12 checkout conversion studies of an observational method yielding a lift estimate very close to that produced by the RCT.
- In **study 12** we did not have access to the data that allowed us to run the "2" and "3" specifications. The RCT lift is 76%. The observational methods we could estimate massively overstated the lift; estimates range from 1231 to 2760%.

Figure 11 summarizes results for the four studies for which there was a conversion pixel on a registration page. Figure 12 summarizes results for the three studies for which there was a

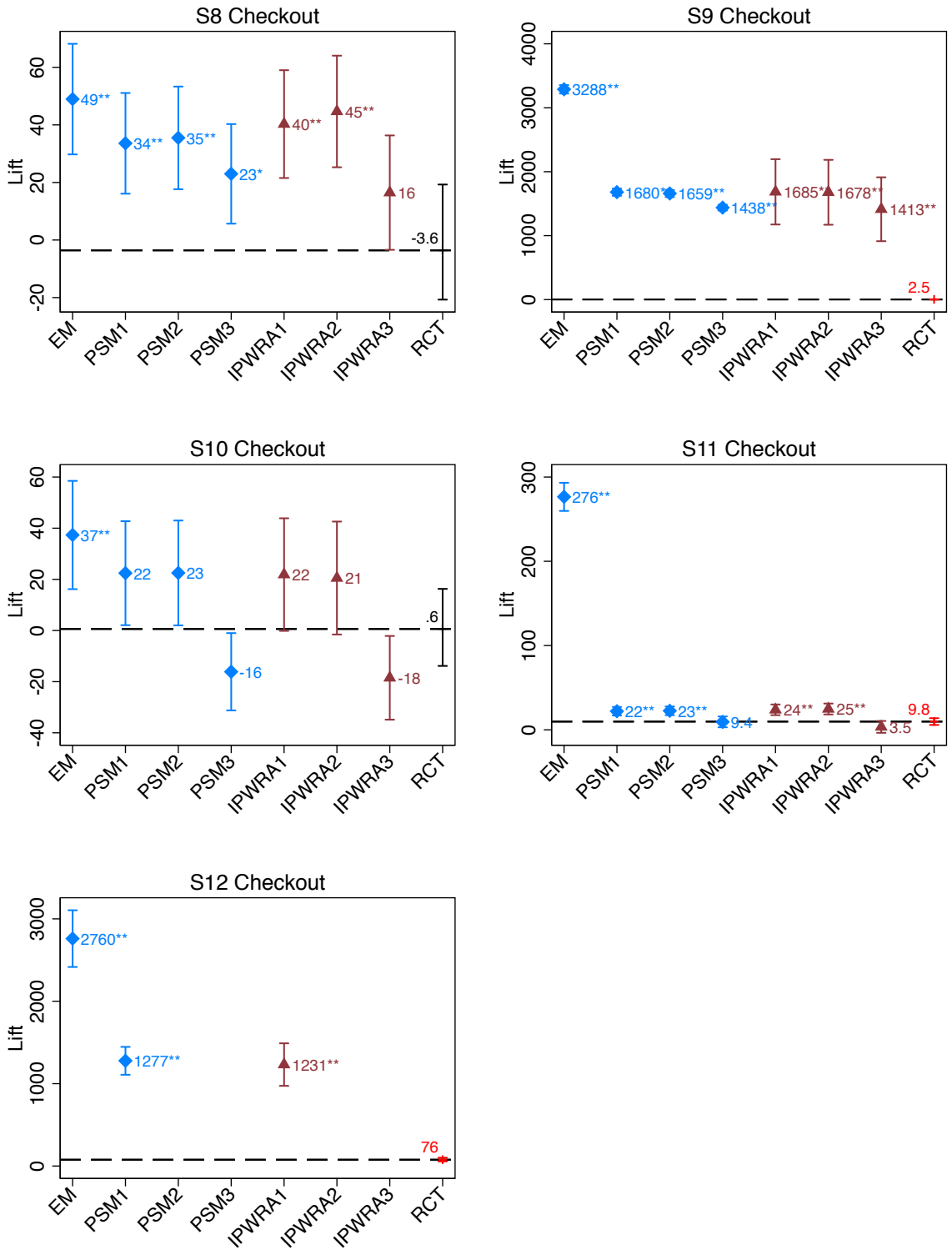
Figure 9: Results for checkout conversion event, studies 1-7



[*] Lift of method significantly different from RCT lift at 5% level; [**] at 1% level.

[RCT Lift in red]: statistically different from zero at 5% level.

Figure 10: Results for checkout conversion event, studies 8-12



[*] Lift of method significantly different from RCT lift at 5% level; [**] at 1% level.

[RCT Lift in red]: statistically different from zero at 5% level.

conversion pixel on a key landing page. The results for these studies vary across studies in how they compare to the RCT results, just as they do for the checkout conversion studies reported in Figures 9 and 10.

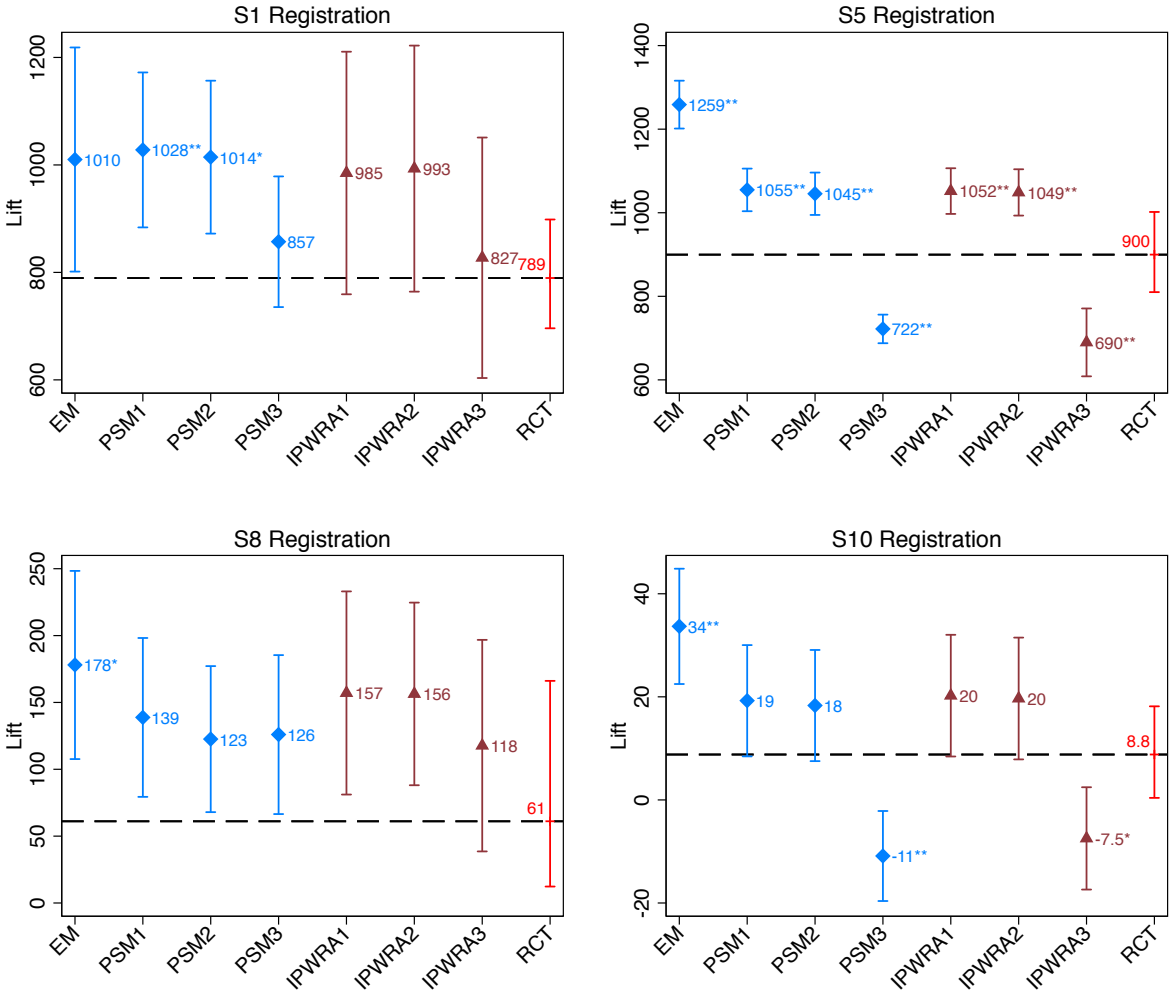
We summarize the performance of different observational approaches using two different metrics. We want to know first how often an observational study fails to capture the truth. Said in a statistically precise way, "For how many of the studies do we reject the hypothesis that the lift of the observational method is equal to the RCT lift?" Table 7 reports the answer to this question. We divide the table by outcome reported in the study (checkout is in the top section of Table 7, followed by registration and page view). The first row of Table 7 tells us that of the 11 studies that tracked checkout conversions, we statistically reject the hypothesis that the exact matching estimate of lift equals the RCT estimate. As we go down the column, the propensity score matching and regression adjustment approaches fare a little better, but for all but one specification, we reject equality with the RCT estimate for half the studies or more.

We would also like to know how different the estimate produced by an observational method is from the RCT estimate. We present the average absolute deviation in percentage points between the observational method estimate of lift and the RCT lift. For example, the RCT lift for study 1 (checkout outcome) is 33%. The EM lift estimate is 117%. Hence the absolute lift deviation is 84 percentage points. For study 2 (checkout outcome) the RCT lift is 0.9%, the EM lift estimate is 535%, and the absolute lift deviation is 534 percentage points. Averaging over all studies, exact matching leads to an average absolute lift deviation of 661 percentage points relative to an average RCT lift of 57% across studies (see the last two columns of the first row of the table.)

As the table shows, inverse probability weighted regression adjustment with the most detailed set of variables (IPWRA3) yields the smallest average absolute lift deviation across all evaluated outcomes. For checkout outcomes, the deviation is large, namely 173 vs. an average RCT lift of 57%. For registration and page view outcomes, however, the average absolute lift deviation is relatively small, namely 80 vs. an average RCT lift of 440%, and 94 vs. an average RCT lift of 744%.

In general, observational methods to a better job of approximating RCT outcomes for registration and page view outcomes than for checkouts. We believe that the reason for this lies in the nature of these outcomes. Since unexposed users (in both treatment and control) are comparatively unlikely to find a registration or landing page on their own, comparing the exposed group in treatment to a subset of the unexposed group in the treatment group (the comparison all observational methods are based on) yields relatively similar outcomes to comparing the exposed group in treatment to the (always unexposed) control group (the comparison the RCT is based on).

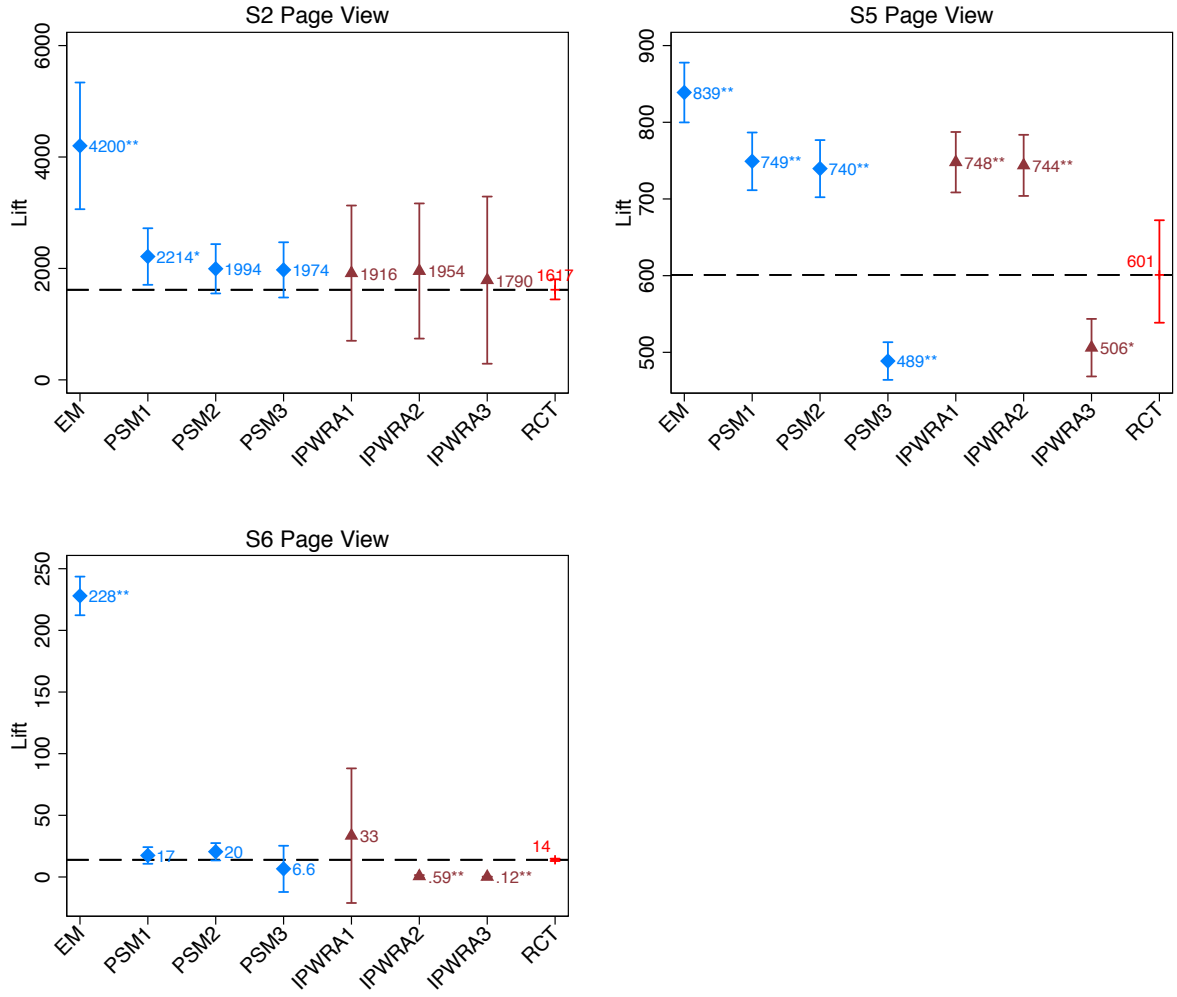
Figure 11: Results for registration conversion event



[*] Lift of method significantly different from RCT lift at 5% level; [**] at 1% level.

[RCT Lift in red]: statistically different from zero at 5% level.

Figure 12: Results for key page view conversion event



[*] Lift of method significantly different from RCT lift at 5% level; [**] at 1% level.

[RCT Lift in red]: statistically different from zero at 5% level.

Table 7: Summary of performance by method for different conversion types

Method	Outcome evaluated	# of studies	# of studies with Lift ≠ RCT Lift*	% of studies with Lift ≠ RCT Lift*	Average absolute Lift deviation from RCT Lift in percentage points	Average RCT Lift in percent
EM	Checkout	11	10	91	661	57
PSM1	Checkout	11	9	82	296	57
PSM2	Checkout	10	8	80	202	57
PSM3	Checkout	10	5	50	184	57
IPWRA1	Checkout	11	8	73	288	57
IPWRA2	Checkout	10	7	70	201	57
IPWRA3	Checkout	10	3	30	173	57
EM	Registration	4	3	75	180	440
PSM1	Registration	4	2	50	120	440
PSM2	Registration	4	2	50	110	440
PSM3	Registration	4	2	50	82	440
IPWRA1	Registration	4	1	25	114	440
IPWRA2	Registration	4	1	25	115	440
IPWRA3	Registration	4	2	50	80	440
EM	Page View	3	3	100	1012	744
PSM1	Page View	3	2	67	250	744
PSM2	Page View	3	1	33	174	744
PSM3	Page View	3	1	33	159	744
IPWRA1	Page View	3	1	33	155	744
IPWRA2	Page View	3	2	67	165	744
IPWRA3	Page View	3	2	67	94	744

* Difference is statistically significant at a 5% level.

8 Conclusion

In this paper we have analyzed whether and when observational methods can reliably substitute for randomized experiments in online advertising measurement. We have done so by using a collection of 12 large-scale advertising RCTs conducted at Facebook. We used the outcomes of these studies to reconstruct different sets of observational methods for measuring ad effectiveness and then compared each of them to the results obtained from the RCT.

Our results showed that observational methods could some times—but not reliably—replicate the result from an RCT. In some cases, one or more of these methods obtained ad lift estimates that were statistically indistinguishable from those of the RCT. However, even the best method produced an average absolute deviation of 173% relative to an average RCT lift of 57% for checkout conversion outcomes, with most of the methods yielding upwardly biased estimates. Results were somewhat better for registration and page view outcomes, where the best methods produced an average absolute deviation of 80% and 94% relative to an average RCT lift of 400% and 744%, respectively.

Our paper has made two contribution. First, we have shown that—in contrast to the belief

in industry that observational methods for ad measurement are "good enough"—the nature of selection in online advertising does not seem to make observational methods a reliable alternative to RCTs for online ad effectiveness measurement. Similar to the critique of RCTs in other fields (Deaton 2010, Hausman 2016), RCTs in online advertising can only test a few advertising strategies relative to the enormous space of possible advertising strategies. This highlights one potential benefit of observational methods, which is that, relative to RCT's, much more data for high-dimensional problems is typically available because the data are generated more easily and by more actors. However, this presumes that observational methods can correct for selection using suitable observables and estimation approaches. Our results suggests that they can not. One caveat in coming to this conclusion is that the performance of the observational methods we study is only as good as the data we have at our disposal. It is possible that better data, for example time-varying user-level data on online activity and generalized shopping behavior, would significantly improve the performance of observational methods. We should note, however, that industry insiders have told us that the data we use in this paper is at par with (and potentially better than) what is normally available to industry researchers.

Our second contribution is to add to the literature on observational vs. experimental approaches to causal measurement. Over the last two decades we have seen enormous improvements in observational methods for causal inference (Imbens and Rubin 2015). In this paper we have analyzed whether the improvements in observational methods for causal inference are sufficient for replicating experimentally generated results in a large industry where such methods are commonly used in practice. We have found they do not—at least with the data we had at our disposal.

Our paper presents a work in progress. The degree to which our analysis is useful depends on the quality of both the data and methods. We are planning to make improvements on both fronts in future versions of this paper. First, we expect to obtain time-varying user-level data on online activity and generalized shopping behavior, which might allow us to better control for activity bias. Second, we plan to implement additional observational methods (such as blocking/regression) and to more carefully examine the overlap in observables between users in the matched exposed and unexposed groups.

References

Abadie, A., and G. Imbens (2015): "Matching on the Estimated Propensity Score," *Working paper, Stanford GSB*.

Abadie, A., and G. W. Imbens (2008): "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76(6), 1537–1557.

Abraham, M. (2008): "The off-line impact of online ads," *Harvard Business Review*, 86(4), 28.

Andrews, D. W. K., and M. Buchinsky (2000): "A three-step method for choosing the number of bootstrap repetitions," *Econometrica*, 68(1), 213–251.

Blake, T., C. Nosko, and S. Tadelis (2015): "BlakeNoskoTadelis2015," *Econometrica*, 83(1), 155–174.

Caliendo, M., and S. Kopeinig (2008): "Some Practical Guidance for the Implementation of Propensity Score Matching," *Journal of Economic Surveys*, 22(1), 31–72.

comScore (2010): "comScore Announces Introduction of AdEffx Smart Control^{TM} Ground-Breaking Methodology for Measuring Digital Advertising Effectiveness," Press Release.

Deaton, A. (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48, 424–455.

Deheji, R., and S. Wahba (2002): "Propensity Score Matching Methods for Non-Experimental Causal Studies," *The Review of Economics and Statistics*, 84(1), 151–161.

DellaVigna, S., and M. Gentzkow (2010): "Persuasion: Empirical Evidence," *Annual Review of Economics*, 2.

Fawcett, T. (2006): "An introduction to ROC analysis," *Pattern Recognition Letters*, 27, 861–874.

Gluck, M. (2011): "Best Practices for Conducting Online Ad Effectiveness Research," Report, Interactive Advertising Bureau.

Goldfarb, A., and C. Tucker (2011a): "Online Display Advertising: Targeting and Obtrusiveness," *Marketing Science*, 30(3), 389–404.

——— (2011b): "Search Engine Advertising: Channel Substitution When Pricing Ads to Context," *Management Science*, 57(3), 458–70.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

Hausman, R. (2016): "The Problem With Evidence-Based Policies," Discussion paper, Harvard University.

Heckman, J. J., H. Ichimura, J. Smith, and P. E. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64(4), 605–654.

——— (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Imbens, G., and D. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 1st edn.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86(1), 4–29.

——— (2015): "Matching Methods in Practice: Three Examples," *The Journal of Human Resources*, 50(2), 373–419.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Imbens, G. W., W. Newey, and G. Ridder (2005): "Mean-Squared-Error Calculations for Average Treatment Effects," Unpublished manuscript, Department of Economics, UC Berkeley.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

Johnson, G. A., R. Lewis, and E. Nubbemeyer (2015a): "The Online Display Ad Effectiveness Funnel and Carry-Over: A Meta-study of Ghost Ad Experiments," Working paper, University of Rochester.

Johnson, G. A., R. Lewis, and D. Reiley (2015): "Location, Location, Location: Repetition and Proximity Increase Advertising Effectiveness," Working paper, University of Rochester.

Johnson, G. A., R. A. Lewis, and E. I. Nubbemeyer (2015b): "Ghost Ads: Improving the Economics of Measuring Ad Effectiveness," *Working paper, Simon Business School*.

Klein, C., and L. Wood (2013): "Cross Platform Sales Impact: Cracking The Code On Single Source," Report, Nielsen Catalina Solutions and Time Inc.

Lalonde, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4), 604–620.

Lavrakas, P. (2010): "An Evaluation of Methods Used to Assess the Effectiveness of Advertising on the Internet," Report, Interactive Advertising Bureau.

Lewis, R., and D. Nguyen (2015): "Display advertising's competitive spillovers to consumer search," *Quantitative Marketing and Economics*, 13(2), 93–115.

Lewis, R., and J. Rao (forthcoming): "The unfavorable economics of measuring the returns to advertising," *Quarterly Journal of Economics*.

Lewis, R., J. Rao, and D. Reiley (2011): "Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimtes of the Effects of Advertising," in *Proceedings of the 20th International Conference on World Wide Web*, pp. 157–66. Association for Computing Machines.

——— (2015): "Measuring the effects of advertising: The digital frontier," in *Economic Analysis of the Digital Economy*, ed. by A. Goldfarb, S. Greenstein, and C. Tucker. University of Chicago Press.

Lewis, R., and D. Reiley (2014): "Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on Yahoo!," *Quantitative Marketing and Economics*, 12(3), 235–266.

Politis, D. N., and J. P. Romano (1994): "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions," *The Annals of Statistics*, 22(4), 2031–2050.

Robins, J., and Y. Ritov (1997): "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine*, 16, 285–319.

Rosenbaum, P., and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrica*, 70, 41–55.

Rosenbaum, P., and D. Rubin (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate The Propensity Score," *American Statistician*, 39, 33–38.

Rubin, D. (1978): "Bayesian inference for causal effects: The role of randomization," *Annals of Statistics*, 6, 34–58.

Rubin, D., and N. Thomas (2000): "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates," *Journal of the American Statistical Association*, 95(450), 573–585.

Rutz, O., and R. Rucklin (2011): "From Generic to Branded: A Model of Spillover in Paid Search Advertising," *Journal of Marketing Research*, 48(1), 87–102.

Sahni, N., and H. Nair (2016): "Native Advertising, Sponsorship Disclosure and Consumer Deception: Evidence from Mobile Search-Ad Experiments," .

Stuart, A., and K. Ord (2010): *Kendall's Advanced Theory of Statistics, Distribution Theory*, vol. 1. Wiley.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

——— (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141, 1281–1301.

ONLINE APPENDIX

Lift Confidence Intervals

Below we have copied equation (7) from section 2 that defines lift:

$$\text{Lift} = \frac{\text{Actual conversion rate} - \text{Counterfactual conversion rate}}{\text{Counterfactual conversion rate}}$$

To facilitate exposition, we rewire the above with some notation:

$$\text{Lift} = \frac{y_e(e) - y_e(u)}{y_e(u)}$$

where $y_e(e)$ is the conversion rate of the exposed users assuming they had actually been exposed and $y_e(u)$ is the conversion rate of exposed users had they instead been unexposed. The former is directly observed in the data whereas the latter requires a model to generate the counterfactual prediction. Next we can rewrite this equation another way, using the fact that the counterfactual conversion rate is the difference between the actual conversion rate and the estimated average treatment effect on the treated (ATT), which is $y_e(u) = y_e(e) - ATT$, and gives us:

$$\begin{aligned}
\text{Lift} &= \frac{y_e(e) - y_e(u)}{y_e(u)} \\
&= \frac{y_e(e) - (y_e(e) - ATT)}{y_e(e) - ATT} \\
&= \frac{ATT}{y_e(e) - ATT}
\end{aligned}$$

To determine the confidence interval on the lift, we require the standard error of the numerator and the denominator. The standard error of the ATT is available in each of the methods we consider. In the denominator, the standard error on $y_e(e)$ is straightforward to calculate because, unlike the ATT, the term does not rely on a model to estimate it. That is, given the set of relevant exposed users, we calculate the standard error on their conversion rates using the usual formula for a standard error. However, the tricky issue is that the numerator and denominator are clearly not independent. This implies we must calculate the covariance between the numerator and denominator to estimate the standard error on the lift. The exception is when we can performing a bootstrap is feasible and the standard error can be calculated from the bootstrapped samples. We discuss our procedures for estimating the standard errors for each method below.

- RCT Lift. Rather than estimating the covariance explicitly, we implement a nonparametric bootstrap to calculate the confidence intervals for the RCT lift estimates. We use the method in Andrews and Buchinsky (2000) to choose a suitable number of bootstrap draws to ensure an accurate estimate of the confidence interval. This approach has the advantage that it automatically integrates uncertainty about $y_e(e)$, the ATT, the share of exposed users, and the ratio statistic.

- IPWRA. We recover the covariance for the estimates through the covariance matrix estimated from the GMM procedure in Stata. This output contains separate estimates of the ATT and

$(y_e(e) - ATT)$, estimates for the standard errors of each term, and the covariance estimate. We can substitute these point estimates for the means, standard errors and covariance into the following approximation (based on Taylor expansions) for the variance of the ratio of two (potentially dependent) random variables:

$$Var\left(\frac{x}{y}\right) \approx \left(\frac{E(x)}{E(y)}\right)^2 \left(\frac{Var(x)}{E(x)^2} + \frac{Var(y)}{E(y)^2} - 2\frac{Cov(x,y)}{E(x)E(y)}\right)$$

The interested reader should refer to Stuart and Ord (2010).

- PSM. The standard errors for the ATT are computed using the methods explained in Abadie and Imbens (2015) to account for the uncertainty in the propensity score estimates. The standard error for the conversion rate of exposed matched users ($y_e(e)$) is calculated directly from the data using the standard formula. However, no formal results exist to estimate the covariance between the ATT and conversion rate of exposed users. Instead, we implement a subsampling procedure (Politis and Romano 1994) to generate multiple estimates of the ATT and the conversion rate of the exposed users, since bootstrapping is invalid in the context of matching prodedures (Abadie and Imbens 2008). We calculate the covariance based on these results and use it to construct the standard error on the lift using the approximation above. In general, the covariance is small enough relative to the standard error of each term that both the quantitative and qualitative conclusions of the various hypothesis tests are unaffected.