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

Brett R. Gordon
Kellogg School of Management
Northwestern University

Florian Zettelmeyer
Kellogg School of Management
Northwestern University and NBER

Neha Bhargava
Facebook

Dan Chapsky
Facebook

April 12, 2018

Abstract

Measuring the causal effects of digital advertising remains challenging despite the availability of granular data. Unobservable factors make exposure endogenous, and advertising's effect on outcomes tends to be small. In principle, these concerns could be addressed using randomized controlled trials (RCTs). In practice, few online ad campaigns rely on RCTs, and instead use observational methods to estimate ad effects. We assess empirically whether the variation in data typically available in the advertising industry enables observational methods to recover the causal effects of online advertising. This analysis is of particular interest because of recent, large improvements in observational methods for causal inference (Imbens and Rubin 2015). Using data from 15 US advertising experiments at Facebook comprising 500 million user-experiment observations and 1.6 billion ad impressions, we contrast the experimental results to those obtained from multiple observational models. The observational methods often fail to produce the same effects as the randomized experiments, even after conditioning on extensive demographic and behavioral variables. We also characterize the incremental explanatory power our data would require to enable observational methods to successfully measure advertising effects. Our findings suggest that commonly used observational approaches based on the data usually available in the industry often fail to accurately measure the true effect of advertising.

Keywords: Digital Advertising, Field Experiments, Causal Inference, Observational Methods, Advertising Measurement.

* To maintain privacy, no data contained personally identifiable information that could identify consumers or advertisers. We thank Daniel Slotwiner, Gabrielle Gibbs, Joseph Davin, Brian d'Alessandro, and Fangfang Tan at Facebook. We are grateful to Garrett Johnson, Randall Lewis, Daniel Zantedeschi, and seminar participants at Bocconi, CKGSB, Columbia, eBay, ESMT, Facebook, FTC, HBS, LBS, Northwestern, QME, Temple, UC Berkeley, UCL, NBER Digitization, NYU Big Data Conference, and ZEW 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 exceeded television ad spending for the first time in 2017.[1] Advertising is a critical funding source for internet content and services (Benady 2016). As advertisers have shifted more of their ad expenditures online, demand has grown for online ad effectiveness measurement: advertisers routinely access granular data that link ad exposures, clicks, page visits, online purchases, and even offline purchases (Bond 2017).

However, even with these data, measuring the causal effect of advertising remains challenging for at least two reasons. First, individual-level outcomes are volatile relative to ad spending per customer, such that advertising explains only a small amount of the variation in outcomes (Lewis and Reiley 2014, Lewis and Rao 2015). Second, even small amounts of advertising endogeneity (e.g., likely buyers are more likely to be exposed to the ad) can severely bias causal estimates of its effectiveness (Lewis, Rao, and Reiley 2011).

In principle, using large-scale randomized controlled trials (RCTs) to evaluate advertising effectiveness could address these concerns.[2] In practice, however, few online ad campaigns rely on RCTs (Lavrakas 2010). Reasons range from the technical difficulty of implementing experimentation in ad-targeting engines to the commonly held view that such experimentation is expensive and often unnecessary relative to alternative methods (Gluck 2011). Thus, many advertisers and leading ad-measurement companies rely on observational methods to estimate advertising's causal effect (Abraham 2008, comScore 2010, Klein and Wood 2013, Berkovich and Wood 2016).

Here, we assess empirically whether the variation in data typically available in the advertising industry enables observational methods to recover the causal effects of online advertising. To do so, we use a collection of 15 large-scale advertising campaigns conducted on Facebook as RCTs in 2015. We use this dataset to implement a variety of matching and regression-based methods and compare their results with those obtained from the RCTs. Earlier work to evaluate such observational models had limited individual-level data and considered a narrow set of models (Lewis, Rao, and Reiley 2011, Blake, Nosko, and Tadelis 2015).

A fundamental assumption underlying observational models is unconfoundedness: conditional on observables, treatment and (potential) outcomes are independent. Whether this assumption is true depends on the data-generating process, and in particular on the requirement that some random variation exists after conditioning on observables. In our context, (quasi-)random variation in exposure has at least three sources: user-level variation in visits to Facebook, variation in Facebook's pacing of ad delivery over a campaign's pre-defined window, and variation due to unrelated advertisers' bids. All three forces induce randomness in the ad auction outcomes. However,

[1] `https://www.recode.net/2017/12/4/16733460/2017-digital-ad-spend-advertising-beat-tv`, accessed on April 7, 2018.

[2] A growing literature focuses on measuring digital ad effectiveness using randomized experiments. See, for example Lewis and Reiley (2014), Johnson, Lewis, and Reiley (2016), Johnson, Lewis, and Reiley (2017), Kalyanam, McAteer, Marek, Hodges, and Lin (2018), Johnson, Lewis, and Nubbemeyer (2017a), Johnson, Lewis, and Nubbemeyer (2017b), Sahni (2015), Sahni and Nair (2016), and Goldfarb and Tucker (2011). See Lewis, Rao, and Reiley (2015) for a recent review.

three mechanisms generate endogenous variation between exposure and conversion outcomes: user-induced endogeneity ("activity bias," Lewis et al. 2011), targeting-induced endogeneity due to the ad system overweighing users who are predicted to convert, and competition-induced endogeneity due to the auction mechanism. For an observational model to recover the causal effect, the data must sufficiently control for the endogenous variation without absorbing too much of the exogenous variation.

Our data possess several key attributes that should facilitate the performance of observational methods. First, we observe an unusually rich set of user-level, user-time-level, and user-time-campaign-level covariates. Second, our campaigns have large sample sizes (from 2 million to 140 million users), giving us both statistical power and means to achieve covariate balance. Third, whereas most advertising data are collected at the level of a web browser cookie, our data are captured at the user level, regardless of the user's device or browser, ensuring our covariates are measured at the same unit of observation as the treatment and outcome.[3] Although our data do not correspond exactly to what an advertiser would be able to observe (either directly or through a third-party measurement vendor), our intention is to approximate the data many advertisers have available to them, with the hope that our data are in fact better.

An analysis of our 15 Facebook campaigns shows a significant difference in the ad effectiveness obtained from RCTs and from observational approaches based on the data variation at our disposal. Generally, the observational methods overestimate ad effectiveness relative to the RCT, although in some cases, they significantly underestimate effectiveness. The bias can be large: in half of our studies, the estimated percentage increase in purchase outcomes is off by a factor of three across all methods.

These findings represent the *first contribution* of our paper, namely, to shed light on whether—as is thought in the industry—observational methods using good individual-level data are "good enough" for ad measurement, or whether even good data prove inadequate to yield reliable estimates of advertising effects. Our results support the latter.

Moreover, our setting is a preview of what might come next in marketing science. The field continues to adopt techniques from data science and large-scale machine learning for many applications, including advertising, pricing, promotions, and inventory optimization. The strong selection effects we observe in digital advertising, driven by high-dimensional targeting algorithms, will likely extend to other fields in the future. Thus, the data requirements necessary to use observational models will continue to grow, increasing the need to develop and integrate experimentation directly into any targeting platform.

One critique of our finding that even good data prove inadequate to yield reliable estimates of

[3] Most advertising data are collected through cookies at the user-device-web-browser level, with two potential consequences. First, users in an experimental control group may inadvertently be simultaneously assigned to the treatment group. Second, advertising exposure across devices may not be fully captured. We avoid both problems because Facebook requires users to log in to Facebook each time they access the service on any device and browser. Therefore, ads are never inadvertently shown to users in the control group, and all ad exposures and outcomes are measured. Lewis and Reiley (2014) also used a sample of logged-in users to match the retailer's existing customers to their Yahoo! profiles.

advertising effects is that we do not observe all the data that Facebook uses to run its advertising platform. Motivated by this possibility, we conducted the following thought experiment: "Assuming 'better' data exist, how much better would that data need to be to eliminate the bias between the observational and RCT estimates?" This analysis, extending work by Rosenbaum and Rubin (1983a) and Ichino, Fabrizia, and Nannicini (2008), begins by simulating an unobservable that eliminates bias in the observational method. Next, we compare the explanatory power of this (simulated) unobservable with the explanatory power of our observables. Our results show that for some studies, we would have to obtain additional covariates that exceed the explanatory power of our full set of observables to recover the RCT estimates. These results represent the *second contribution* of our paper, which is to characterize the nature of the unobservable needed to use observational methods successfully to estimate ad effectiveness.

The *third contribution* of our paper is to the literature on observational versus 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 concludes that " many of the econometric procedures do not replicate the experimentally determined results" (p. 604). Since then, we have seen significant 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 experimental results. In the job-training setting in LaLonde (1986), observational methods needed to adjust for the fact that the characteristics of trainees differed from those of a comparison group drawn from the population. Because of targeting, the endogeneity problems associated with digital advertising are potentially more severe: advertising exposure is determined by a sophisticated machine-learning algorithm using detailed data on individual user behavior. We explore whether the improvements in observational methods for causal inference, paired with large sample, individual-level data, are sufficient to replicate experimental results in a large industry that relies on such methods.

We are not the first to attempt to estimate the performance of observational methods in gauging digital advertising effectiveness.[4] Lewis, Rao, and Reiley (2011) is the first paper to compare RCT estimates with results obtained using observational methods (comparing exposed versus unexposed users and regression). They faced the challenge of finding a valid control group of unexposed users: their experiment exposed 95% of all US-based traffic to the focal ad, leading them to use a matched sample of unexposed international users. Blake, Nosko, and Tadelis (2015) documents that non-experimental measurement can lead to highly suboptimal spending decisions for online search ads. However, in contrast to our paper, Blake, Nosko, and Tadelis (2015) use a difference-in-differences approach based on randomization at the level of 210 media markets as the experimental benchmark and therefore cannot implement individual-level causal inference methods.

This paper proceeds as follows. We first describe the experimental design of the 15 advertising

[4] Beyond digital advertising, other work assesses the effectiveness of marketing messages using both observational and experimental methods in the context of voter mobilization (Arceneaux, Gerber, and Green 2010) and water-usage reduction (Ferraro and Miranda 2014, Ferraro and Miranda 2017).

RCTs we analyze: how advertising works at Facebook, how Facebook implements RCTs, and what determines advertising exposure. In section 3, we introduce the potential-outcomes notation now standard for causal inference and relate it to the design of our RCTs. In section 4, we explain the set of observational methods we analyze. Section 5 presents the data generated by the 15 RCTs. Section 6 discusses identification and estimation issues and presents diagnostics. Section 7 shows the results for one example ad campaign in detail and summarizes findings for all remaining ad campaigns. Section 8 assesses the role of unobservables in reducing bias. Section 9 offers concluding remarks.

2 Experimental Design

Here we describe how Facebook conducts advertising campaign experiments. Facebook enables advertisers to run experiments to measure marketing-campaign effectiveness, test out different marketing tactics, and make more informed budgeting decisions.[5] We define the central measurement question, discuss how users are assigned to the test group, and highlight the endogenous sources of exposure to an ad.

2.1 Advertising on Facebook

We focus exclusively on campaigns in which the advertiser had a particular "direct response" outcome in mind, for example, to increase sales of a new product.[6] The industry refers to these as "conversion outcomes." In each study, the advertiser measured conversion outcomes using a piece of Facebook-provided code ("conversion pixel") embedded on the advertiser's web pages, indicating whether a user visited that page.[7] Different placement of the pixels can measure different conversion outcomes. A conversion pixel embedded on a checkout-confirmation page, for example, measures a purchase outcome. A conversion pixel on a registration-confirmation page measures a registration outcome, and so on. These pixels allow the advertiser (and Facebook) to record conversions for users in both the control and test group and do not require the user to click on the ad to measure conversion outcomes.

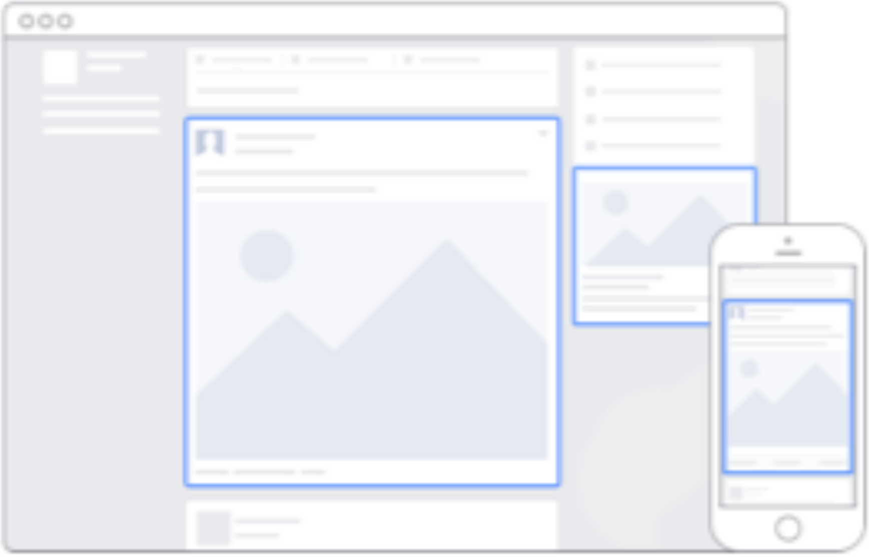
Facebook's ability to track users via a "single-user login" across devices and sessions represents a significant measurement advantage over more common cookie-based approaches. First, this approach helps ensure the integrity of the random assignment mechanism because a user's assignment can be maintained persistently throughout the campaign and prevents control users from being inadvertently shown an ad. Second, Facebook can associate all exposures and conversions across

[5] Facebook refers to these ad tests as "conversion lift" tests (`https://www.facebook.com/business/a/conversion-lift`, accessed on April 7, 2018.). Facebook provides this experimental platform as a free service to qualifying advertisers.

[6] We excluded brand-building campaigns in which outcomes are measured through consumer surveys.

[7] A "conversion pixel" refers to two types of pixels used by Facebook. One is traditionally called a "conversion pixel," and the other is known as a "Facebook pixel." The studies analyzed in this paper use both types, and they are equivalent for our purposes (`https://www.facebook.com/business/help/460491677335370`, accessed on April 7, 2018).

Figure 1: Facebook desktop and mobile-ad placement



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

devices and sessions with a particular user. Such cross-device tracking is critical because users are frequently exposed to advertising on a mobile device but might subsequently convert on a tablet or computer.

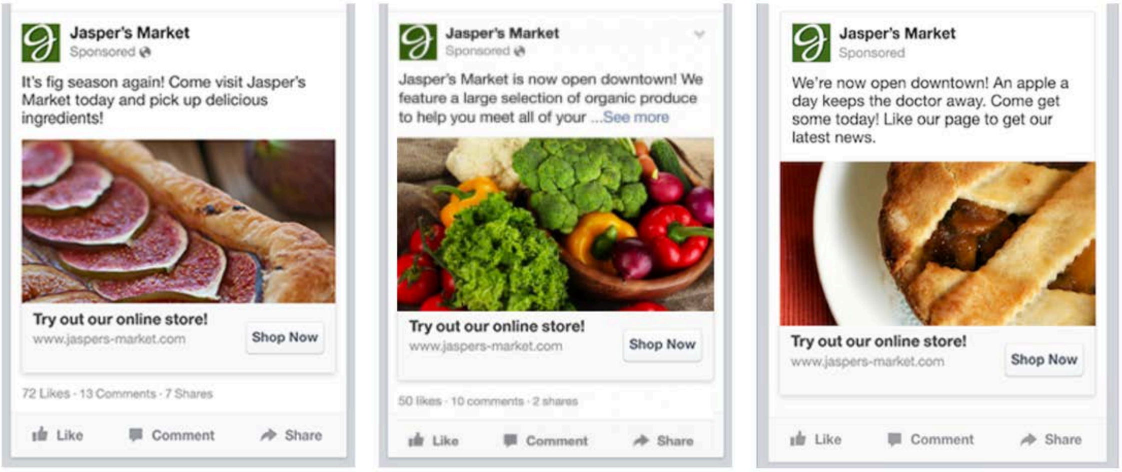
Figure 1 displays where a Facebook user accessing the site from a desktop/laptop or mobile device might see ads. In the middle is the "News Feed," where new stories appear with content as the user scrolls down or the site automatically refreshes. Ads appear as tiles in the News Feed, with a smaller portion served to the right of the page. News Feed ads are an example of "native advertising" because they appear interlaced with organic content. On mobile devices, only the News Feed is visible; no ads appear on the right side. The rate at which Facebook serves ads in the News Feed is carefully managed at the site level, independent of any ad experiment.

An advertising campaign is a collection of related advertisements ("creatives") served during the campaign period. A campaign may have multiple associated ads, as Figure 2 illustrates for Jasper's Market, a fictitious advertiser. Although imagery and text vary across ads in a campaign, the overall message is generally consistent. We evaluate the effect of the whole campaign, not the effects of specific ads.

As with most online advertising, each impression is the result of an underlying auction. 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.[8] The auction plays a role in the experiment's implementation and in generating endogenous variation in exposures, both of which are discussed in the following sections.

[8] Additional factors beyond the advertiser's bid determine the actual ranking. For more information, see `https://www.facebook.com/business/help/430291176997542`, accessed on April 7, 2018.

Figure 2: Example of three display ads for one campaign



Jasper's Market
Sponsored

It's fig season again! Come visit Jasper's Market today and pick up delicious ingredients!

Try out our online store!
www.jaspers-market.com

72 Likes · 13 Comments · 7 Shares

Jasper's Market
Sponsored

Jasper's Market is now open downtown! We feature a large selection of organic produce to help you meet all of your ...See more

Try out our online store!
www.jaspers-market.com

50 likes · 10 comments · 2 shares

Jasper's Market
Sponsored

We're now open downtown! An apple a day keeps the doctor away. Come get some today! Like our page to get our latest news.

Try out our online store!
www.jaspers-market.com

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

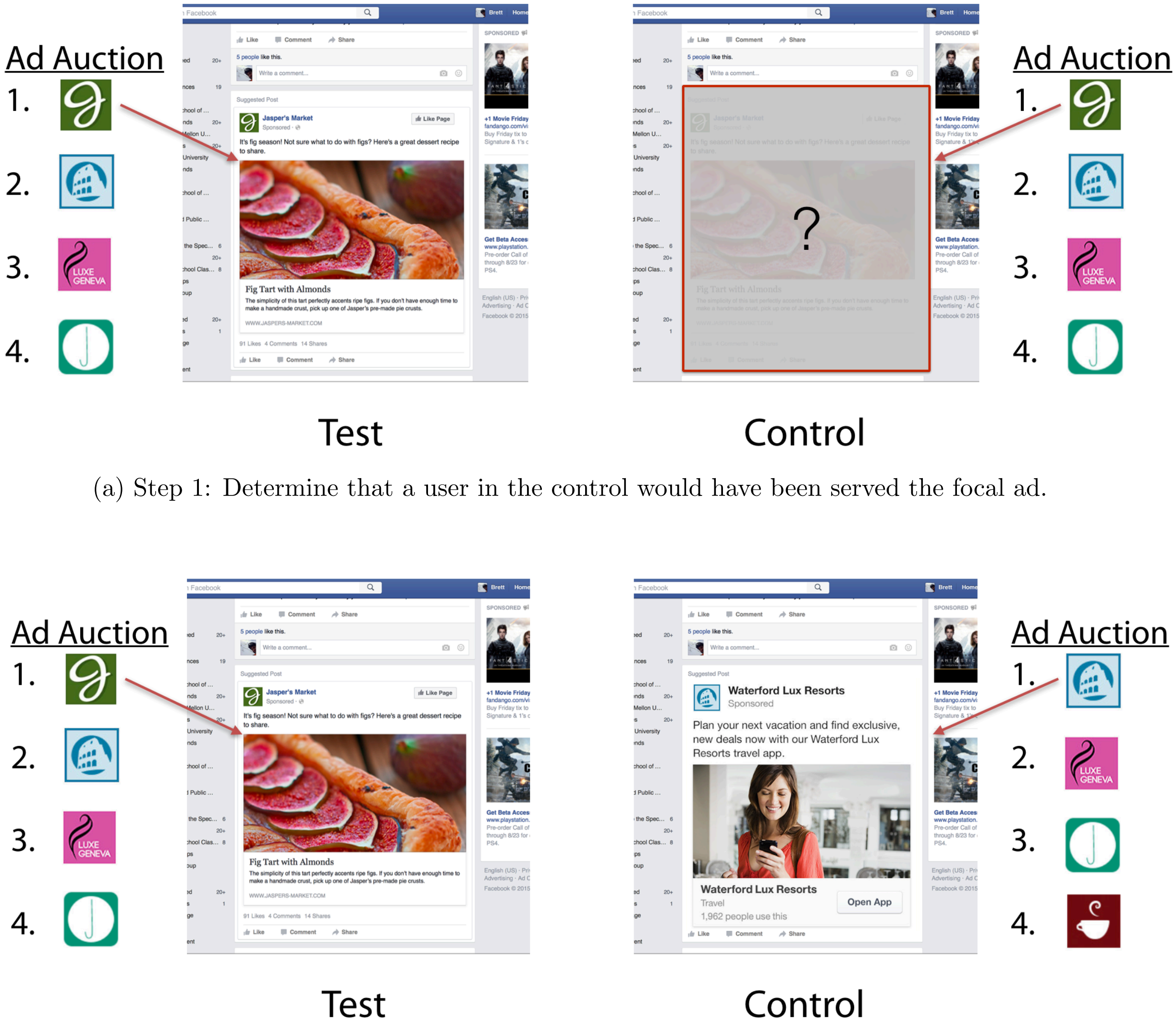
2.2 Experimental Implementation

An experiment begins with the advertiser deciding which consumers to target with a marketing campaign, such as all women between 18 and 54. These targeting rules define the relevant set of users in the study. Each user is randomly assigned to the control or test group based on a proportion selected by the advertiser, in consultation with Facebook. Control-group members are never exposed to campaign ads during the study; those in the test group are eligible to see the campaign's ads. Facebook avoids contaminating the control group with exposed users, due to its single-user login feature. Whether test-group users are ultimately exposed to the ads depends on factors such as whether the user accessed Facebook during the study period (we discuss these factors and their implications in the next subsection). Thus, we observe three user groups: control-unexposed, test-unexposed, and test-exposed.

Next, we consider what ads the control group should be shown in place of the advertiser's campaign. This choice defines the counterfactual of interest. To evaluate campaign effectiveness, an advertiser requires the control condition to estimate the outcomes that would have occurred without the campaign. Thus, the control-condition ads should be the ads that *would have been* served if the advertiser's campaign had not been run on Facebook.

We illustrate this process using a hypothetical, stylized example in Figure 3. Consider two users in the test and control groups. Suppose that at a given moment, Jasper's Market wins the auction to display an impression for the test-group user, as seen in Figure 3a. Imagine the control-group user, who occupies a parallel world to that of 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 appearing. As Figure 3b shows, instead the auction's second-place ad is served to the control user because that user would have won the auction if the focal ad had not existed.

Figure 3: Determination of control ads in Facebook experiments


Ad Auction
1.
2.
3.
4.
LUXE GENEVA
Jasper's Market
Sponsored
It's fig season! Not sure what to do with figs? Here's a great dessert recipe to share.
Fig Tart with Almonds
The simplicity of this tart perfectly accents ripe figs. If you don't have enough time to make a handmade crust, pick up one of Jasper's pre-made pie crusts.
WWW.JASPERS-MARKET.COM
?
Test
Control
Waterford Lux Resorts
Sponsored
Plan your next vacation and find exclusive, new deals now with our Waterford Lux Resorts travel app.
Waterford Lux Resorts
Travel
1,962 people use this
Open App

(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.

We must emphasize that this experimental mechanism is relevant only for users in the control group, because it substitutes the second-place ad for the focal ad if the focal ad wins the auction for what they see. In the example, Waterford Lux Resorts is the "control ad" shown to the control user. At another instance when Jasper's Market would have won the auction, a different advertiser might occupy the second-place rank. Thus, rather than a single control ad, users in the control condition are shown the full distribution of ads they would have seen if the advertiser's campaign had not run.

This approach relies on the auction mechanism's stability to the removal of the focal ad. That is, the second-place ad is the same whether the focal advertiser participated in the auction or not. 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 and occur over relatively short periods. Furthermore, Facebook's scale makes gauging other campaigns' scope or targeting objectives hard for advertisers.

As with any experiment, this one yields an estimate of the campaign's average treatment effect, conditional on all market conditions—such as marketing activities the advertiser conducts in other channels (e.g., search, TV) and its competitors' activities. The estimated lift the experiment yields may not generalize to similar future campaigns if market conditions change. If advertising effects are nonlinear across media, the experiment measures something akin to the average *net* effect of the campaign given the distribution of non-Facebook advertising exposures across the sample.

2.3 Determinants of Advertising Exposure

In the experiments, compliance is perfect for users in the control group, who are never shown campaign ads. However, compliance is one-sided in the test group, where exposure (receipt of treatment) is an endogenous outcome that depends on factors related to the user, platform, and advertisers. These factors generate systematic differences (i.e., selection bias) between exposed and unexposed test-group users. Three features of online advertising environments in general make the selection bias of exposure particularly significant.

User-induced endogeneity

The first mechanism that drives selection was coined "activity bias" when first identified by Lewis, Rao, and Reiley (2011). In our context, activity bias arises because a user must visit Facebook during the campaign to be exposed. If conversion is a purely digital outcome (e.g., online purchase, registration), exposed users will be more likely to convert merely because they happened to be online during the campaign. For example, a vacationing target-group user may be less likely to visit Facebook and therefore miss the ad campaign. What leads to endogeneity is that the user may also be less likely to engage in any online activities, such as online purchasing. Thus, the conversion rate of the unexposed group provides a biased estimate of the conversion rate of the exposed group had it not been exposed.

Targeting-induced endogeneity

The targeting criteria for the campaign determines the pool of potential users who may be assigned to the test or control group at the start of the campaign. Although these criteria do not change once the campaign begins, modern advertising delivery systems optimize who are shown ads. Multiple targeting objectives exist, with the most common being maximizing the number of impressions, click-through rate, or purchase. As a campaign progresses, the delivery system learns which types of users are most likely to meet the objective, and gradually the system starts to favor showing ads to users it expects are most likely to meet the objective. To implement this, the delivery system upweights or downweights the auction bids of different types of users within the target group. As a result, conditional on the advertiser's bid, the probability of exposure increases or decreases for different users.

Assessing ad effectiveness by comparing exposed versus unexposed consumers will, therefore, overstate the effectiveness of advertising because exposed users were specifically chosen based on their higher conversion rates. In general, this mechanism will lead to upwardly-biased ad effects, but there are cases where the bias could run in the opposite direction. One example is if the ad campaign is set to optimize for clicks but the advertiser still tracks purchases. Users who are more likely to click on an ad (so-called "clicky users") may also be less likely to purchase the product.

Note that the implementation of this system at Facebook does not invalidate experimentation, because the upweighting or downweighting of bids is applied equally to users in the test and control group. Some users in the test group may become more likely to see the ad if the system observes similar users converting in the early stages of the campaign. The key point is that the same process occurs for users in the control group: the focal ad will receive more weight in the auction for these users and might win the auction more frequently—except that, for members of the control group, the focal ad is replaced "at the last moment" by the runner up and is thus never shown. As a result, the control group remains a valid counterfactual for outcomes in the treatment group, even under ad-targeting optimization.

Competition-induced endogeneity

Ads are delivered if the advertiser wins the auction for a particular impression. Winning the auction implies the advertiser outbid other advertisers competing for the same impression. Therefore, an advertiser's ads are more likely to be shown to users the advertiser values highly, most often those with a higher expected conversion probability. Even if an advertiser's actions do not produce any selection bias, the advertiser can nevertheless end up with selection bias in exposures because of what another advertiser does. For example, if, during the campaign period, another advertiser bids high on 18-54-year-old women who are also mothers, the likelihood that mothers will not be exposed to the focal campaign is higher. A case that could lead to downward bias is when other firms sell complementary products and target the same users as a focal advertiser. If these firms win impressions at the expense of the focal advertiser, and obtain some conversions as a result, the resulting set of unexposed users may now be more likely to buy the focal firm's product.

In the RCT, we address potential selection bias by leveraging the random-assignment mechanism and information on whether a user receives treatment. For the observational models, we discard the randomized control group and address the selection bias by relying solely on the treatment status and observables in the test group.

3 Analysis of the RCT

We use the potential-outcomes notation now standard in the literature on experimental and non-experimental program evaluation. Our exposition in this section and the next draws heavily on material in 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. 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. We observe a set of covariates $X_i \in \mathbb{X} \subset \mathbb{R}^P$ for each user that are unaffected by the experiment. We do not index any variable by a study-specific subscript, because all analysis takes place within a study.

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.

Valid inference requires several standard assumptions. First, a user can receive only one version of the treatment, and a user's treatment assignment does not interfere with another user's outcomes. This pair of conditions is commonly known as the Stable Unit Treatment Value Assumption (SUTVA), a term coined in Rubin (1978). Our setting likely satisfies both conditions. Facebook's ability to track individuals prevents the platform from inadvertently showing the wrong treatment to a given user. Non-interference could be violated if, for example, users in the test group share ads with users in the control group. However, users are unaware of both the existence of the experiment and their assignment status. Moreover, if test users shared ads with control users on Facebook, we would be able to observe those impressions.[9]

[9] If test users showed control users the ads, the treatment-effect estimates would be conservative because it might inflate the conversion rate in the control group.

The second assumption is that assignment to treatment is random, or 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, the received W_i is not necessarily random, due to one-sided non-compliance. This assumption is untestable because we do not observe all potential outcomes and treatments. We have performed a variety of randomization checks on each study and failed to find any evidence against proper randomization.

In principle, we could focus on the relationship between the random assignment Z_i and outcome Y_i, ignoring information in W_i. Such an intent-to-treat (ITT) analysis only requires the two assumptions above.[10]

However, the primary goal of this paper is to compare treatment effects from RCTs with those obtained from observational methods; thus, the treatment effects must be inherently comparable. Because we exclude the control group from our analysis using the observational methods, we cannot produce ITT estimates using both approaches. Instead, all our analysis compares the average treatment effect on the treated (ATT)—the effect of the ads on users who are actually exposed to ads. Depending on their goals, managers evaluating ad effectiveness might be interested in the ITT, ATT, or both. We focus on the ATT to facilitate comparison with the results from the observational models.

The ATT requires one more assumption: an exclusion restriction,

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

such that assignment affects a user's outcome only through receipt of the treatment. Because users are unaware of their assignment status, only exposure should affect outcomes. This permits Z_i to serve as an instrumental variable (IV) to recover the ATT.

3.2 Causal Effects in the RCT

Given the assumptions, the ITT effect of assignment on outcomes compares across random-assignment status,

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

with the sample analog being

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

As noted earlier, our focus is on the ATT,

$$\text{ATT} = \mathbb{E}\left[Y(1, W(1)) - Y(0, W(0)) | W(1) = 1\right] \ . \tag{4}$$

[10] When we usually interpret an ITT, it is always conditional on the entire treatment (e.g., a specific ad delivered on a particular day and time on a specific TV network) and who is targeted with the treatment. In the context of online advertising, the "entire treatment" includes the advertising platform, including its ad-optimization system. Hence, the ITT should be interpreted as conditional on the platform's ad-optimization system.

Note that the ATT is inherently conditional on the set of users who end up being exposed (or treated) in a particular experiment. As different experiments target individuals using different X's, the interpretation of the ATT varies across experiments. Imbens and Angrist (1994) show the ATT can be expressed in an IV framework, relying on the exclusion restriction. The ATT is the ITT effect on the outcome, divided by the ITT effect on the receipt of treatment:

$$\tau = \frac{\text{ITT}_Y}{\text{ITT}_W} = \frac{\mathbb{E}[Y(1,W(1))] - \mathbb{E}[Y(0,W(0))]}{\mathbb{E}[W(1)] - \mathbb{E}[W(0)]} \tag{5}$$

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 $\text{ITT}_W = \mathbb{E}[W(1)]$, or the proportion in the test group who take up the treatment. In summary, we go from ITT to ATT by using the (exogenous) treatment assignment Z as an instrument for (endogenous) exposure W.

An intuitive way to derive the relationship between the ITT and the ATT is to decompose the ITT outcome effect for the entire sample as the weighted average of the effects for two groups of users: *compliers* and *noncompliers*. Compliers are users assigned to the test condition who receive the treatment, $W_i(1) = 1$, and noncompliers are users assigned to the test condition who do not receive the treatment, $W_i(1) = 0$. The overall ITT effect can be expressed as

$$\text{ITT}_Y = \text{ITT}_{Y,co} \cdot \pi_{co} + \text{ITT}_{Y,nc} \cdot (1 - \pi_{co}), \tag{6}$$

where $\pi_{co} = \mathbb{E}[W(1)]$ is the share of compliers. The exclusion restriction assumes unexposed users have the same outcomes, regardless of whether they were in treatment or control, $Y_i(1,0) = Y_i(0,0)$. This 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 ITT effect divided by the share of compliers,

$$\tau \equiv ATT \equiv \text{ITT}_{Y,co} = \frac{\text{ITT}_Y}{\pi_{co}} \ . \tag{7}$$

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). Imbens and Angrist (1994) refer to this quantity as the local average treatment effect (LATE) and demonstrate its relationship to IV with heterogeneous treatment effects. If the sample contains no "always-takers" and no "defiers," which is true in our experimental design with one-sided non-compliance, the LATE is equal to the ATT.

3.3 Lift

To help summarize outcomes across advertising studies, we report most results in terms of *lift*, the incremental conversion rate among treated users expressed as a percentage:

$$\begin{aligned} \tau_\ell &= \frac{\Delta\text{Conversion rate due to ads in the treated group}}{\text{Conversion rate of the treated group if they had } \textit{not} \text{ been treated}} \\ &= \frac{\tau}{\mathbb{E}[Y^{obs}|Z=1, W^{obs}=1] - \tau} \end{aligned} \tag{8}$$

The denominator is the estimated conversion rate of the treated group if they had not actually been treated. Reporting the lift facilitates comparison of advertising effects across studies because it normalizes the results according to the treated group's baseline conversion rate, which can vary significantly with study characteristics (e.g., advertiser's identity, outcome of interest). One downside of using lift is that differences between methods can seem large when the treated group's baseline conversion rate is small. Other papers have compared advertising effectiveness across campaigns by calculating advertising ROI (Lewis and Rao 2015), but we lack the data on profit margins from sales to calculate ROI.[11]

4 Observational Approaches

Here we present the observational methods we compare with estimates from the RCT. The following thought experiment motivates our analysis. Rather than conducting an RCT, an advertiser (or a third party acting on the advertiser's behalf) followed customary practice by choosing a target sample and making all users eligible to see the ad. Although all users in the sample are eligible to see the ad, only a subsample is eventually exposed. To estimate the treatment effect, the advertiser compares the outcomes in the exposed group with the outcomes in the unexposed group. This approach is equivalent to creating a test sample without a control group held out.

We employ a set of methods that impose various degrees of structure to recover the treatment effects. Our goal is twofold: to ensure we cover the range of observational methods commonly used by academics and practitioners and to understand the extent to which more sophisticated techniques are potentially better at reducing the bias of estimates compared with RCT estimates. The observational methods we use rely on a combination of approaches: matching, stratification, and regression.[12]

Both academics and practitioners rely on the methods we implement. In the context of measuring advertising effectiveness, matching methods appear in a variety of related academic work, such as comparing the efficacy of internet and TV ads for brand building (Draganska, Hartmann, and Stanglein 2014), measuring the effects of firm-generated social media on customer metrics (Kumar, Bezawada, Rishika, Janakiraman, and Kannan 2016), assessing whether access to digital video recorders (DVRs) affects sales of advertised products (Bronnenberg, Dubé, and Mela 2010), the effectiveness of pharmaceutical detailing (Rubin and Waterman 2007), the impact of mutual fund name changes on subsequent investment inflows (Cooper, Gulen, and Rau 2005), and evaluating alcohol advertising targeted at adolescents (Ellickson, Collins, Hambarsoomians, and McCaffrey 2005). Industry-measurement vendors, such as comScore, Nielsen, and Nielsen Catalina Solutions, all rely on matching and regression methods to evaluate various marketing programs

[11] Although ROI is a monotone transformation of lift, measuring the ROI in addition to lift would be useful because managerial decisions may rely on cutoff rules that involve ROI.

[12] Researchers have recently developed more sophisticated methods for estimating causal effects (Imai and Ratkovic 2014), including those that blend insights from operations research (Zubizarreta 2012, Zubizarreta 2015) and machine learning (Athey, Imbens, and Wager forthcoming). We leave to future work to explore how these methods perform in recovering experimental ad effects.

(Abraham 2008, comScore 2010, Klein and Wood 2013, Berkovich and Wood 2016). Although obtaining detailed information on the exact nature of these vendors' implementations is difficult, discussions with several industry experts and public case studies confirm these methods are in active use.[13]

4.1 Definitions and Assumptions

To mimic the 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 redefining

$$Y_i(W_i) \equiv Y_i(Z_i = 1, W_i) \ . \tag{9}$$

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

$$Y_i^{obs} \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0 \\ Y_i(1) & \text{if } W_i = 1 \end{cases} \tag{10}$$

The ATT obtained using observational method m is

$$\tau^m = \mathbb{E}\left[Y(1) - Y(0)|W = 1\right] \tag{11}$$

and the lift is

$$\tau_\ell^m = \frac{\tau^m}{\mathbb{E}[Y^{obs}|W = 1] - \tau^m} \ . \tag{12}$$

As before, the denominator in the lift is an estimate of the conversion rate of exposed users if they had been unexposed.

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

$$\tau^{eu} = \mathbb{E}[Y(1) - Y(0)|X] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \tag{13}$$

with corresponding lift of

$$\tau_\ell^{eu} = \frac{\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]}{\mathbb{E}[Y(0)]} \ .$$

Estimates based on comparing exposed and unexposed users serve as a naive baseline.

In reality, W_i is unlikely to be independent of X_i, especially in the world of online advertising. The effect τ^{eu} will contain selection bias due to the relationship between user characteristics, treatment, and outcomes. Observational methods attempt to correct for this bias. Accomplishing this, beyond SUTVA, requires two additional assumptions. The first is unconfoundedness:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \ \mid \ X_i \ , \tag{14}$$

[13] For specific examples, see the case studies at `https://www.ncsolutions.com/pandora-pop-tarts-groove-to-the-tune-of-3x-roas/` and `https://www.ncsolutions.com/frozen-entrees/`, accessed on April 7, 2018.

which states that, conditional on X_i, potential outcomes are independent of treatment status. Alternatively, this assumption posits that no unobserved characteristics of individuals associated with the treatment and potential outcomes exist. This particular assumption is considered the most controversial and is untestable without an experiment.

The second assumption is overlap, which requires a positive probability of receiving treatment for all values of the observables, such that

$$0 < \Pr(W_i = 1|X_i) < 1, \quad \forall X_i \in \mathbb{X} \ .$$

Overlap can be assessed before and after adjustments are made to each group. Rosenbaum and Rubin (1983b) refer to the combination of unconfoundedness and overlap assumptions as strong ignorability.

The conditional probability of treatment given observables X_i is known as the propensity score,

$$e(x) \equiv \Pr(W_i = 1|X_i = x) \tag{15}$$

Under strong ignorability, Rosenbaum and Rubin (1983b) establish that treatment assignment and the potential outcomes are independent, conditional on the propensity score,

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \ \mid \ e(X_i) \ . \tag{16}$$

Given two individuals with the same propensity scores, exposure status is as good as random. Thus, adjusting for the propensity score eliminates the bias associated with differences in the observables between treated and untreated individuals. This result is central to many of the observational methods widely employed in the literature.

4.2 Observational methods

Exact matching (EM)

Matching is an intuitive method for estimating treatment effects under strong ignorability. To estimate the ATT, matching methods find untreated individuals similar to the treated individuals and use the outcomes from the untreated individuals to impute the missing potential outcomes for the treated individuals. The difference between the actual outcome and the imputed potential outcome is an estimate of the individual-level treatment effect, and averaging over treated individuals yields the ATT. This calculation highlights an appealing aspect of matching methods: they do not assume a particular form for an outcome model.

The simplest approach is to compare treated and untreated individuals who match exactly on a set of observables $X^{em} \subset \mathbb{X}$. To estimate the treatment effect, for each exposed user i, we find the set of unexposed users $|\mathcal{M}_i^c|$ for whom $X_i^{em} = X_j^{em}, j \in \mathcal{M}_i^c$. For an exposed user, we observe $Y_i^{obs} = Y_i(1)$ and require an estimate of the potential outcome $Y_i(0)$. An estimate of this potential outcome is

$$\widehat{Y_i(0)} = \frac{1}{\mathcal{M}_i^c} \sum_{j \in \mathcal{M}_i^c} Y_j^{obs} \ . \tag{17}$$

The exact matching estimator for the ATT is

$$\widehat{\tau}^{em} = \frac{1}{N_e} \sum_{i=1}^{N_e} W_i \left(Y_i(1) - \widehat{Y_i(0)} \right) \ , \tag{18}$$

where $N_e = \sum_i^N W_i$ is the number of exposed users (in the test group).

Propensity score matching (PSM)

Exact matching is only feasible using a small set of discrete observables. Generalizing to all the observables requires a similarity metric to compare treated and untreated individuals. One may match on all the observables X_i using a distance metric, such as Mahalanobis distance (Rosenbaum and Rubin 1985). But this metric may not work well with a large number of covariates (Gu and Rosenbaum 1993) and can be computationally demanding.

To overcome this limitation, perhaps the most common matching approach is based on the propensity score (Dehejia and Wahba 2002, Caliendo and Kopeinig 2008, Stuart 2010). Let $e(x;\phi)$ denote the model for the propensity score parameterized by ϕ, the estimation of which we discuss in section 6.2. We match on the (estimated) log-odds ratio

$$\ell(x;\phi) = \ln \left(\frac{e(x;\phi)}{1 - e(x;\phi)} \right) \ .$$

This transformation linearizes values on the unit interval and can improve estimation (Rubin 2001).

To estimate the treatment effect, we find the M unexposed users with the closest propensity scores to each exposed user. Matching is done with replacement because it can reduce bias, does not depend on the sort order of the data, and is less computationally burdensome. Let $m^c_{i,k}$ be the index of the (unexposed) control user that is the k^{th} closest to exposed user i based on $|e(x_{m^c_{i,k}};\phi) - e(x_i;\phi)|$. The set $\mathcal{M}^c_i = \{m^c_{i,1}, m^c_{i,2}, \ldots, m^c_{i,M}\}$ contains the M closest observations for user i. For an exposed user, we observe $Y^{obs}_i = Y_i(1)$ and require an estimate of the potential outcome $Y_i(0)$. An estimate of this potential outcome is

$$\widehat{Y_i(0)} = \frac{1}{M} \sum_{j \in \mathcal{M}^c_i} Y^{obs}_j \ . \tag{19}$$

The propensity score matching estimator for the ATT is

$$\widehat{\tau}^{psm} = \frac{1}{N_e} \sum_{i=1}^{N_e} W_i \left(Y_i(1) - \widehat{Y_i(0)} \right) \ . \tag{20}$$

Stratification (STRAT)

The computational burden of matching on the propensity score can be further reduced by stratification on the estimated propensity score (also known as subclassification or blocking). After estimating the propensity score, the sample is divided into strata (or blocks) such that within each stratum, the estimated propensity scores are approximately constant.

Begin by partitioning the range of the linearized propensity scores into J intervals of $[b_{j-1}, b_j)$, for $j = 1, \ldots, J$. Let B_{ij} be an indicator that user i is contained in stratum j,

$$B_{ij} = 1 \cdot \{b_{j-1} < \ell(x_i, \phi) \leq b_j\} \tag{21}$$

Each stratum contains $N_{wj} = \sum_{i=1}^{N} 1 \cdot \{W_i = w\} B_{ij}$ observations with treatment w. The ATT within a stratum is estimated as

$$\widehat{\tau}_j^{strat} = \frac{1}{N_{1j}} \sum_{i=1}^{N} W_i B_{ij} Y_i - \frac{1}{N_{0j}} \sum_{i=1}^{N} (1 - W_i) B_{ij} Y_i \ . \tag{22}$$

The overall ATT is the weighted average of the within-strata estimates, with weights corresponding to the fraction of treated users in the stratum relative to all treated users,

$$\widehat{\tau}^{strat} = \sum_{j=1}^{J} \frac{N_{1j}}{N_1} \cdot \widehat{\tau}_j^{strat} \tag{23}$$

One task that remains is to determine how to create the strata and how many strata to create. Many researchers follow the advice of Cochran (1968) and set $J = 5$ with equal-sized strata. However, Eckles and Bakshy (2017) suggest setting J such that the number of strata increases with the sample size. We follow the approach proposed in Imbens and Rubin (2015), which uses the variation in the propensity scores to determine the number of strata and their boundaries. In brief, the method recursively splits the data at the median propensity score if the two resulting strata have significantly different average propensity scores. Starting with the full sample, this process continues until the t-statistic comparing two potential splits is below some threshold or if the new stratum falls below a minimum sample size. One appealing aspect of this method is that more (narrower) strata will be created in ranges of the data with greater variation in propensity scores, precisely where having more strata helps ensure the within-stratum variation in propensity scores is minimal.

Regression adjustment (RA)

Whereas exact matching on observables, propensity score matching, and stratification do not rely on an outcome model, another class of methods relies on regression to predict the relationship between treatment and outcomes. Perhaps the simplest approach to estimating the causal effect of advertising is a linear regression with covariates,

$$Y_i^{obs} = \alpha + \beta' X_i + \tau^{reg} W_i + \varepsilon_i \ , \tag{24}$$

where τ^{reg} is the ATT assuming strong ignorability. More generally, we want to estimate the conditional expectation

$$\mu_w(x) = \mathbb{E}[Y^{obs} | W = w, X = x] \ . \tag{25}$$

Separate models could be estimated for each treatment level. Given our focus on the ATT, we estimate only $\mu_0(x)$ to predict counterfactual outcomes for the treated users. The most common approach is a linear model of the form $\mu_w(X_i; \beta_w) = \beta_w' X_i$, with flexible functions of X_i. Given some estimator $\mu_0(X_i; \hat{\beta}_0)$, the regression-adjustment (RA) estimate for the ATT is obtained through

$$\widehat{\tau}^{ra} = \frac{1}{N_e} \sum_{i=1}^{N} W_i [Y_i^{obs} - \mu_0(X_i; \hat{\beta}_0)] \ . \tag{26}$$

Note the accuracy of this method depends on how well the covariate distribution for untreated users overlaps the covariate distribution for treated users. If the treated users have significantly different observables compared to untreated users, $\mu_0(X_i; \hat{\beta}_0)$ relies heavily on extrapolation, which is likely to produce biased estimates of the treatment effect in equation (26).

Inverse-probability-weighted regression adjustment (IPWRA)

A variant of the RA estimator incorporates information in the propensity scores, borrowing from the insights found in inverse-probability-weighted estimators (Hirano, Imbens, and Ridder 2003). The estimated propensity scores are used 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), which means the estimator is consistent even if one of the underlying models—either the propensity model or the outcome model—turns out to be misspecified.

The inverse-probability-weighed regression adjustment (IPWRA) model estimates the exposure and outcome models simultaneously:

$$\min_{\{\phi,\beta\}} \sum_{i=1}^{N} W_i \left[\frac{(Y_i - \mu_0(X_i; \beta_0))^2}{1 - e(X_i; \phi)} \right]$$

Given the estimate $\hat{\beta}_0$ from the outcome model, equation (26) is once again used to calculate the treatment effect, $\widehat{\tau}^{ipwra}$. In practice, the exposure model, outcome model, and ATT are estimated simultaneously using two-step GMM to obtain efficient estimates and robust standard errors (Wooldridge 2007).

Stratification and Regression (STRATREG)

One problem with regression estimators, even those that weigh by the inverse propensity scores, is that treatment effects can be sensitive to differences in the covariate distributions for the treated and untreated groups. If these distributions differ, these estimators rely heavily on extrapolation.

A particularly flexible approach, advocated by Imbens (2015) and Eckles and Bakshy (2017), is to combine regression with stratification on the estimated propensity score. After estimating the propensity score, the sample is divided into strata with approximately constant estimated propensity scores. Regression on the outcome is used within each stratum to estimate the treatment effect and to correct for any remaining imbalance. The idea is that the covariate distribution within a stratum should be relatively balanced, so the within-stratum regression is less prone to extrapolate.

Stratification follows the recursive procedure outlined after equation (23), with a regression within each strata j to estimate the strata-specific ATT:

$$Y_i = \alpha_j + \tau_j^{stratreg} \cdot W_i + \beta_j' X_i + \varepsilon_i. \tag{27}$$

As in equation (23), this method produces a set of J estimates that can be averaged appropriately to calculate the ATT:

$$\widehat{\tau}^{stratreg} = \sum_{j=1}^{J} \frac{N_{1j}}{N_T} \cdot \tau_j^{stratreg} \quad . \tag{28}$$

4.3 Alternative Methods and Discussion

The goal of each observational method we have discussed is to find and isolate the random variation that exists in the data, while conditioning on the endogenous variation. The latter is accomplished by matching on covariates (directly or via a propensity score), by controlling for covariates in an outcome model, or both.

A critique of these observational methods is that sophisticated ad-targeting systems aim for ad exposure that is deterministic and based on a machine-learning algorithm. In the limit, such ad-targeting systems would completely eliminate any random variation in exposure, in which case, the observational methods we have discussed in section 4.2 would fail. As an example, consider propensity scoring. If we observed the exact data and structure used by the ad-targeting systems, the propensity score distribution would collapse to discrete masses at 0 and 1. This is not surprising, because a deterministic exposure system implies that common support in observables between treated and untreated observations cannot exist. As a result, any matching system would fail, as would any regression approach that requires common support on observables.

If ad-targeting systems were completely deterministic, identification of causal effects would have to rely on alternative observational methods, for example, regression discontinuity (RD). If the ad-targeting rules were known, an RD design would identify users whose observables are very similar but ended up triggering a different exposure decision by the ad-targeting system. In practice, implementing such an RD approach would require extensive collaboration with the advertising platform, because the advertiser would need to know the full data and structure used by the ad-targeting system. Given that advertisers avoid RCTs partially because RCTs require the collaboration of the platform, RD-type observational methods would unlikely be more popular. Moreover, RD-type observational methods are unlikely to overcome the problem that some platforms cannot implement RCTs: if a platform had the sophistication to run an RD design, it would probably also have the sophistication to implement RCTs.

As of now, ad-targeting systems have not eliminated all exogenous reasons a given person would be exposed to an ad campaign whereas a probabilistically equivalent person would not. As we discuss in detail in section 6.1, in our context, quasi-random variation in exposure has at least three sources: user-level variation in visits to Facebook, variation in Facebook's pacing of ad delivery over the campaign's pre-defined window, and variation in the remaining campaign budget. As a result, the observational methods we have discussed in section 4.2 need not fail. However, as ad-targeting systems become more sophisticated, such failure is increasingly likely.

5 Data

The 15 advertising studies analyzed in this paper were chosen by two of its authors (Gordon and Zettelmeyer) based on criteria to make them suitable for comparing common ad-effectiveness methodologies: conducted after January 2015, when Facebook first made the experimentation platform available to sufficiently large advertisers; minimum sample size of 1 million users; business-relevant conversion tracking in place; no retargeting campaign by the advertiser; and no significant sharing of ads between users. The window during which we obtained studies for this paper was from January to September 2015. Although the sample of studies is not representative of all Facebook advertising (nor is it intended to be), it covers a varied set of verticals (retail, financial services, e-commerce, telecom, and tech), represents a range of sample sizes, and contains a mix of test/control splits. All studies were US-based RCTs, and we restrict attention to users age 18 and older.

Table 1 provides summary statistics for each study. The studies range in size, with the smallest containing around 2 million users and the largest about 140 million, and with a mix of test/control splits. The studies also differed by the conversion outcome(s) the advertiser measured. In all but one study, the advertiser placed a conversion pixel on the checkout-confirmation page to measure whether a Facebook user purchased from the advertiser. In five 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 "page view").

Table 2 provides information on the variables we observe. For most of the observational models, we implement a sequence of specifications corresponding to the grouping of covariates. The first two groups of variables are at the user level but are time- and study-invariant. The third group is indexed by user and time but not by study. The fourth group is at the user, time, and study level. We believe the third and fourth groups of covariates should especially help us account for activity bias in the estimation of treatment effects. The variable sets are

1. **(FB Variables)** The first specification includes variable set 1 from Table 2, which are common Facebook variables such as age, gender, how long users have been on Facebook, how many Facebook friends they have, their phone OS, and other characteristics.

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

3. **(User-Activity Variables)** In addition to the variables in 2, we incorporate data on a user's overall level of activity on Facebook. Specifically, for each user and device type (desktop,

Table 1: Summary statistics for all studies

Study	Vertical	Observations	Test	Control	Impressions	Clicks	Conversions	Outcomes*
1	Retail	2,427,494	50%	50%	39,167,679	45,401	8,767	C, R
2	Finan. serv.	86,183,523	85%	15%	577,005,340	247,122	95,305	C, P
3	E-commerce	4,672,112	50%	50%	7,655,089	48,005	61,273	C
4	Retail	25,553,093	70%	30%	14,261,207	474,341	4,935	C
5	E-commerce	18,486,000	50%	50%	7,334,636	89,649	226,817	C, R, P
6	Telecom	141,254,650	75%	25%	590,377,329	5,914,424	867,033	P
7	Retail	67,398,350	17%	83%	61,248,021	139,471	127,976	C
8	E-commerce	8,333,319	50%	50%	2,250,984	204,688	4,102	C, R
9	E-commerce	71,068,955	75%	25%	35,197,874	222,050	113,531	C
10	Tech	1,955,375	60%	40%	2,943,890	22,390	7,625	C, R
11	E-commerce	13,339,044	50%	50%	11,633,187	106,534	225,241	C
12	Retail	5,566,367	50%	50%	10,070,742	54,423	215,227	C
13	E-commerce	3,716,015	77%	23%	2,121,967	22,305	7,518	C, R
14	E-commerce	86,766,019	80%	20%	36,814,315	471,501	15,722	C
15	Retail	9,753,847	50%	50%	8,750,270	19,365	76,177	C

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

mobile, or other), the raw activity level is measured as the total number of ad impressions served to that user in the week before the start of any given study. We measure the total number of ad impressions across all Facebook campaigns that were running in that week—not just the campaigns in our sample. This approach captures not only how long a user stays on Facebook, but also how much the user scrolls through items in his or her news feed. Our data transform this raw measure into deciles that describe where, for each device, a user ranks in the distribution of all users. We include a full set of dummy variables across deciles and devices to allow for the greatest flexibility of different specifications.

4. **(Match Score)** In addition to the variables in 3, we add a composite metric of Facebook data that summarizes thousands of behavioral variables and is a machine-learning-based metric Facebook uses to construct target audiences similar to consumers an advertiser has identified as desirable.[14] For each study, this metric represents a measure of the similarity between exposed users and all other users from a machine-learning model with thousands of features. Including this variable, and functions of it, in estimating our propensity score allows us to condition on a summary statistic for data beyond which we had direct access and to move beyond concerns that a more flexible propensity-score model might change the results.

To check whether the randomization of the RCTs was implemented correctly, we compared means across test and control for each study and variable, resulting in 1,251 p-values. Of these, 10% are below 0.10, 4% 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 suggests they are and, indeed, a Kolmogorov-Smirnov test fails to reject that the p-values are uniformly distributed on the unit interval (p-value=0.4). We have also

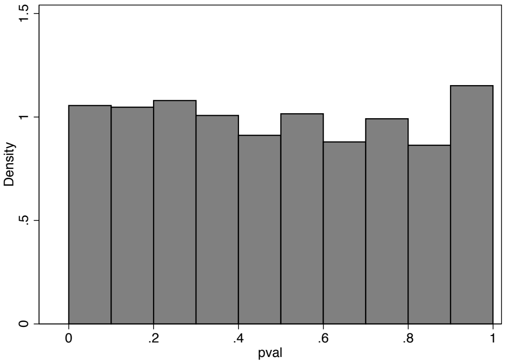
[14] See `https://www.facebook.com/business/help/164749007013531`, accessed on April 7, 2018.

Table 2: Description of variables

Set	Variable	Description	Source
1	age	Age of user	FB
1	gender	1 = female, 0 = otherwise	FB
1	rel_status	married, engaged, in relationship, single, other	FB
1	FB age	Days since user joined FB	FB
1	friends	# of friends	FB
1	num_initiated	# of friend requests sent	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	tablet_OS	operating system of primary tablet (if exists)	FB
1	region	region of user's residence	FB
1	zip	zip code 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	% working in STEM	ACS
2	pctprofessional	% working in professional jobs	ACS
2	pcthealth	% working in health industry	ACS
2	pctprotective	% working in protective 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 administration	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 industry	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	mobile_activity	decile of users' FB activity on mobile devices	FB
3	desktop_activity	decile of users' FB activity on desktop devices	FB
3	other_activity	decile of users' FB activity on other devices	FB
4	match_score	Composite variable of FB data	FB

Notes: First three rows are self-reported by the users. Region and zip code are determined by geolocation. ACS data are from 2010.

Figure 4: Distribution of p-values across all studies



been unable to find evidence that particular variables might be more likely to exhibit imbalance. Thus, we find no evidence that the randomization was implemented improperly.

6 Identification and Estimation

In this section, we discuss the sources of exogenous variation in the data on which the observational methods rely, how we estimate propensity scores and conduct statistical inference, and provide evidence of covariate balance. We follow the best practices detailed in Imbens (2015) for using matching or propensity score methods.

6.1 Identification

In the context of the observational data, which only rely on the test group, highlighting the sources of (quasi-)random variation on which the observational models rely is useful. The goal of each observational method is to find and isolate the random variation that exists, while conditioning on the endogenous variation. Our data contain at least three sources of random variation.

First, the advertising platform at Facebook generates plausibly exogenous variation through the pacing of ad delivery.[15] At the start of a campaign, an advertiser sets a budget and campaign length. The pacing system determines how an advertiser's budget is spent, with the most common goal being to deliver ads smoothly over the course of the campaign. Suppose an advertiser runs a campaign with a budget of $100,000 over four weeks. After one week, the platform observes that $50,000 has already been spent, such that the campaign might end prematurely by exhausting its budget. To avert this outcome, the system will downweight the advertiser's bids in the impression auctions to slow down delivery. The pacing system continuously attempts to learn the optimal bid adjustments, which vary depending on the type of ad, target audience, time of day, and other factors, in order to satisfy the campaign's goal. This implies that ad impressions for a given campaign always contain some variation that is plausibly exogenous to potential user outcomes.

[15] See `https://developers.facebook.com/docs/marketing-api/pacing`, accessed on April 7, 2018.

Second, the pacing is determined by an advertiser's budget and the budgets of all other advertisers competing for the same target audience. Some advertisers' bidding preferences for a particular audience of users may be orthogonal to a focal advertiser's conversion outcome. For instance, a luxury automaker and a yogurt manufacturer may both value the same segment of consumers, but it is hard to imagine how one firm's outcomes could be related to the other firm's ad bids. The implication is that the budgets and bidding strategies of other advertisers can affect the advertising delivery for the focal advertiser in such a way that is likely independent of the focal advertiser's outcomes.

Third, quasi-random behavior is present in the timing of users' visits to Facebook. The timing of a user's visit throughout the day or week is likely influenced by a plethora of random factors specific to that user (e.g., local weather, work schedule, just missing a subway, etc.).

These mechanisms generate exogenous variation in exposure across users and time within a campaign. However, the three sources of endogenous selection into exposure discussed in section 2.3—user, targeting, and competitive—generate confounding variation. Under unconfoundedness, the assumption is that the observational models will rely on the observables to control for the endogenous variation in the data while retaining some of the exogenous variation. Each method controls for this endogenous variation using slightly different parametric forms.

6.2 Estimation and Inference

The propensity score plays a central role in all but one of the observational models. To be consistent with most applications, we model the propensity score using a logistic regression:

$$e(x;\phi) = \frac{\exp(x'\phi)}{1+\exp(x'\phi)} \ .$$

To obtain a sufficiently flexible specification, we consider numerous functions of the covariates for inclusion in the logistic regression. When possible, we convert integer-valued variables into a full set of dummies (e.g., one dummy for each age). We generate interactions and higher-order terms, both within and across the four variable groups, between both dummies and continuous covariates. This approach leads to a large number of covariates, many of which likely have low predictive power and thus might produce low precision in propensity score estimates.

To address this issue, we apply a variant of the LASSO (Tibshirani 1996) developed in Belloni, Chen, Chernozhukov, and Hansen (2012) to estimate the propensity score. This method provides an iterative, data-dependent technique to select the LASSO penalty parameter and to retain a subset of variables for prediction. For methods with an outcome model (RA, IPWRA, STRATREG), we also apply the LASSO to predict Y_i using all the variables, retaining the union of variable sets between the treatment and outcomes models for estimation.[16]

[16] More sophisticated specifications exist, including nonparametric models (Hirano, Imbens, and Ridder 2003), methods from machine learning (McCaffrey, Ridgeway, and Morral 2004, Westreich, Lessler, and Funk 2010), and, more recently, deep learning models (Pham and Shen 2017). Our goal is to choose a reasonable approach that generates estimates similar to other reasonable methods. We explored other techniques, and found they did not produce significantly different treatment effects.

We re-estimate the logistic regression with the subset of variables identified above and apply a simple trimming rule to improve overlap in the covariate distributions. Following Imbens (2015), we trim observations with $e(x;\hat{\phi}) < 0.05$ and $e(x;\hat{\phi}) > 0.95$ and re-estimate the propensity score using the trimmed data. The resulting propensity scores are the values used for treatment effects estimation.

Our analysis faces two challenges regarding proper statistical inference. First, using ATT lift for inference is complicated because it is a ratio. The standard error of the lift's numerator, the ATT, is available in each of the methods we consider. In the denominator, the standard error of the outcome Y_i for exposed users is straightforward to calculate because, unlike the ATT, the term does not rely on a model and so it can be estimated using the usual formula for the standard error of the mean of a Bernoulli random variable. However, because the numerator and denominator are clearly not independent, we must calculate the covariance between them to estimate the standard error on the lift. A second complication is that we wish to conduct hypothesis tests comparing the RCT ATT lift τ_ℓ, defined in equation (8), with the lift obtained from each observational method, τ_ℓ^m. Because the estimates are obtained from the same sample, we must account for the covariance between the estimates when calculating a t-statistic:

$$t = \frac{\tau_\ell - \tau_\ell^m}{\sqrt{Var(\tau_\ell) + Var(\tau_\ell^m) - 2Cov\left(\tau_\ell, \tau_\ell^m\right)}} \tag{29}$$

Signing the direction of this correlation is difficult; thus, knowing the direction of the bias if we were to ignore this term is hard.

We rely on the bootstrap to address both challenges. First, we draw a sample of observations with replacement from the complete RCT and estimate the ATT τ and lift τ_ℓ. Next, we drop the control group and estimate the treatment effects using an observational model m to produce τ^m and τ_ℓ^m. We use the bootstrapped samples to calculate standard errors and confidence intervals for each estimate. In addition, we compute $Cov\left(\tau_\ell, \tau_\ell^m\right)$ to evaluate the t-statistic above.[17]

To summarize, we follow these steps for a given observational model m:

Step 1: Variable selection. Apply the modified Lasso of Belloni, Chen, Chernozhukov, and Hansen (2012) to predict the treatment W_i, producing $\tilde{X}^W \subset X$. If model m includes an outcome model, also apply the modified LASSO to predict Y_i, producing $\tilde{X}^Y \subset X$. Retain the variables $\tilde{X} = \tilde{X}^W \bigcup \tilde{X}^Y$.

Step 2: Analysis using the Bootstrap. For $s = 1, 2, \ldots, S$, draw a sample of N users with replacement from the complete experiment. For each bootstrap replication s:

(a) Estimate the RCT ATT τ and lift τ_ℓ.

[17] One exception to the above concerns propensity score matching. Although Abadie and Imbens (2008) show the bootstrap is invalid for matching procedures, they note that modifications to the bootstrap, such as subsampling (Politis and Romano 1994) and the M-out-of-N bootstrap (Bickel, Gotze, and van Zweet 1997), are valid inferential techniques for matching estimators. Given this, we implement a subsampling procedure to estimate the ATT lift.

(b) Discard the control group.

(c) Trimming: estimate the propensity score $e(x, \hat{\phi})$ using $x \in \tilde{X}$, remove observations where $e(x, \hat{\phi}) < 0.05$ or $e(x, \hat{\phi}) > 0.95$, and re-estimate the propensity score using the trimmed sample.

(d) Use observational model m and the trimmed data to estimate τ^m and lift τ^m_ℓ.

Step 3: Inference. Calculate standard errors and confidence intervals using the bootstrap samples of $(\tau, \tau^m, \tau_\ell, \tau^m_\ell)$. We report bias-corrected standard errors using $S = 2000$.

6.3 Assessing Balance

The key assumption for all the observational methods is unconfoundedness, which implies treatment is independent of potential outcomes after conditioning on observables. Rosenbaum and Rubin (1983b) show that unconfoundedness conditional on the observables implies unconfoundedness conditional on the propensity score. This result is useful because matching on the scalar propensity score is easier than matching on all observables.

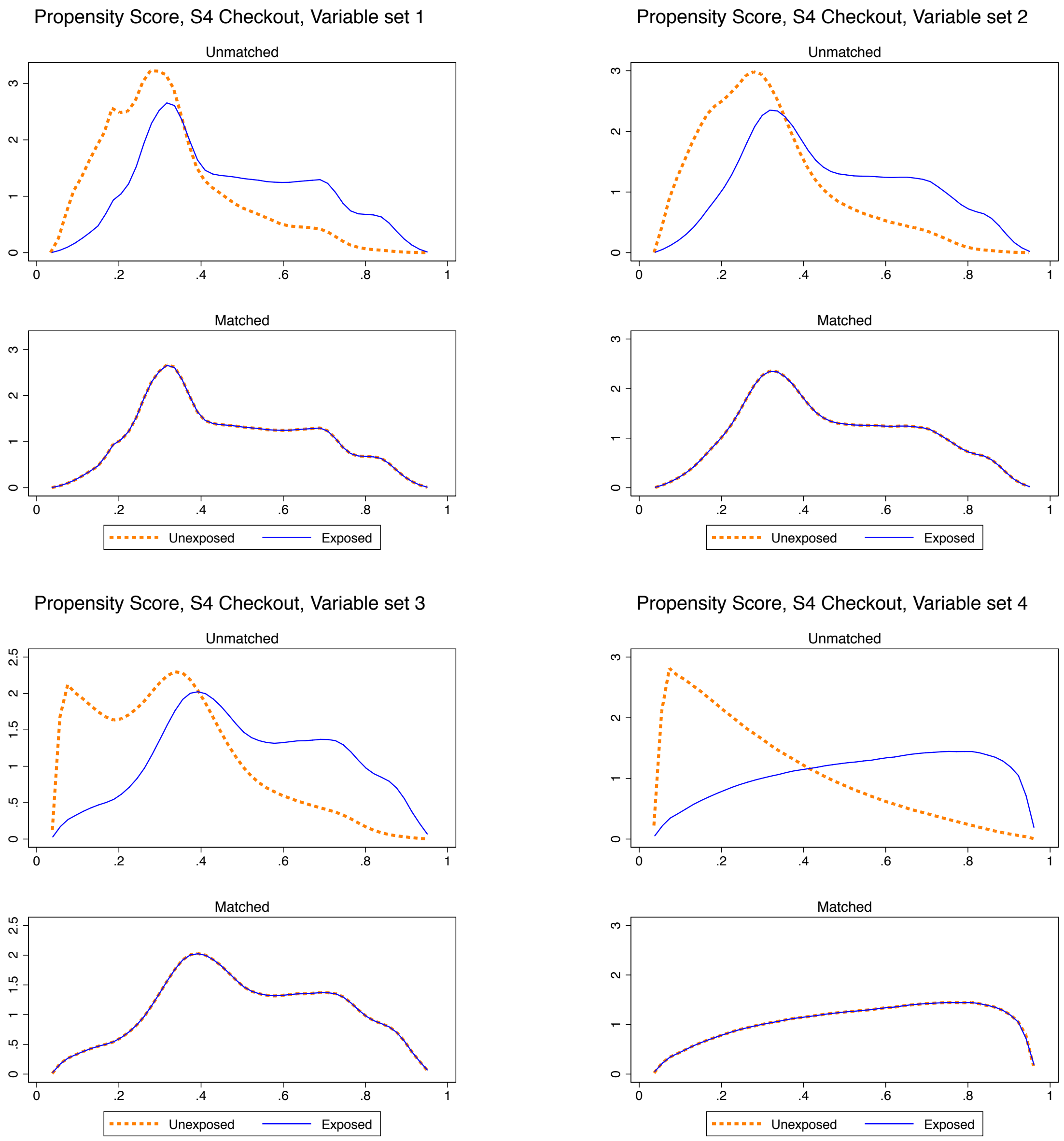
However, because unconfoundedness is fundamentally untestable, researchers have developed strategies to understand whether unconfoundedness might be plausible in any given empirical setting. In methods that utilize propensity scores, a key requirement is that the distribution of propensity scores be balanced across exposed and unexposed users after matching. If such balance is achieved, the hope is that the underlying distribution of observables will also be balanced.

We check both these requirements in all the ad studies. To assess the balance of the propensity scores, we inspect the histograms of the estimated propensity scores by treatment group.[18] Next, we examine standardized differences in covariates between the exposed and unexposed users, before and after matching.

Continuing with Study 4 as our example, Figure 5 presents the density of the estimated propensity scores by treatment status, before and after matching on the propensity scores. The four plots depict the densities obtained from estimating the propensity scores using the same sequence of covariate groups in section 5 (variable sets 1-4). As we add additional covariates, the propensity score densities across exposed and unexposed groups exhibit greater separation, illustrating the predictive power of the covariates. Note the support of the densities is from 0.05 to 0.95 due to trimming. In each case, the unmatched densities share significant overlap but vary considerably over the range. Matching balances the propensity score densities across treatment status, so well in fact that the matched lines overlap perfectly in the plots. Little bias appears to remain in the difference of the propensity scores between the exposed and unexposed group.

[18] We are following the advice of Imbens and Wooldridge (2009), who emphasize that "a major concern in applying methods under the assumption of unconfoundedness is a lack of overlap in the covariate distributions. In fact, once one is committed to the unconfoundedness assumption, this may well be the main problem facing the analyst [...] a direct way of assessing the overlap in covariate distributions is to inspect histograms of the estimated propensity score by treatment status" (page 43).

Figure 5: Study 4 density of estimated propensity scores by treatment and pre/post matching



Detailed graphs with the results of all remaining studies start on page A-1. We present the density of estimated propensity scores by treatment status, using the complete set of observables (variable set 4) to estimate the propensity scores. In each study, we successfully balance the distribution of the propensity score.

Our ability to balance the propensity score distribution is perhaps unsurprising. The sheer size of our studies ensures we have a large pool of unexposed users to match to exposed users, even in studies in which a majority of users are exposed. In such cases, matching with replacement is necessary to match all exposed users.

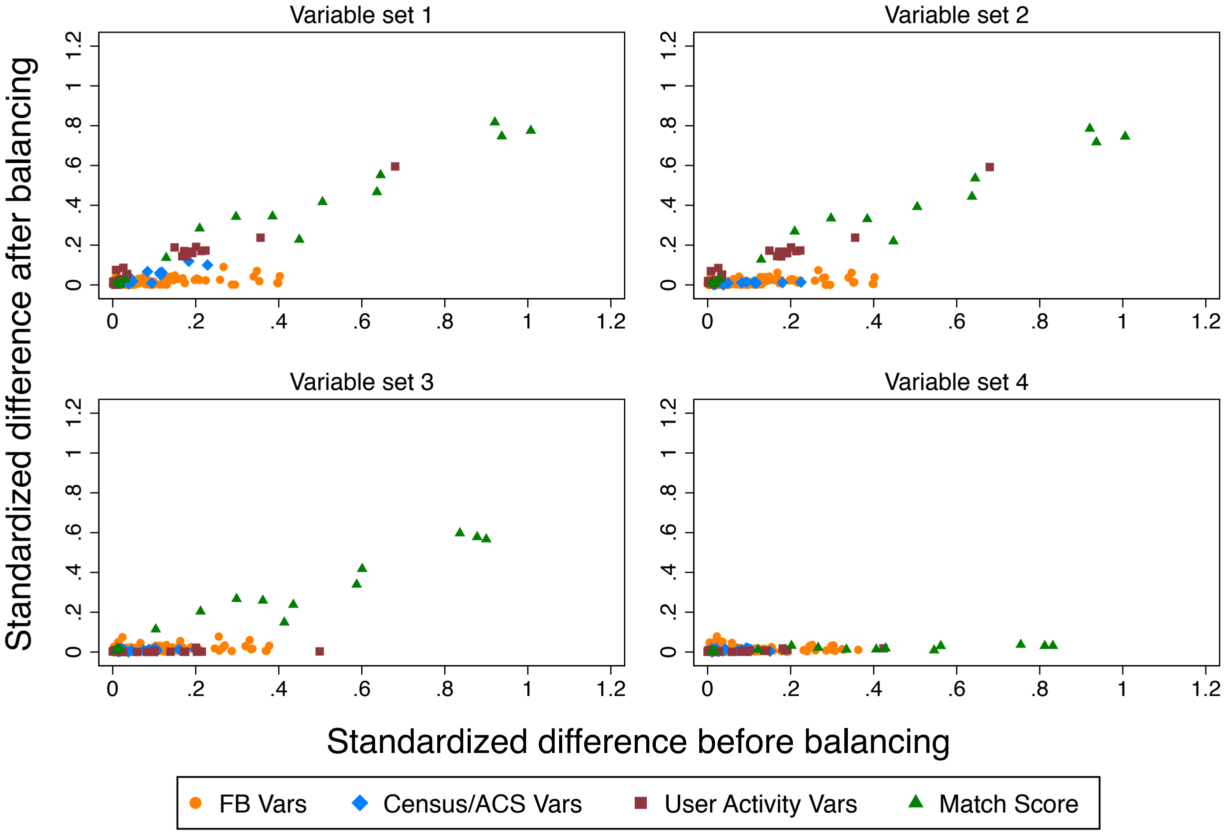
Given that we achieve balance on the propensity score distributions, how well does this balance the actual observables? Following the recommendations of Gelman and Hill (2007) and Imbens and Wooldridge (2009), we examine the standardized differences in covariate means between exposed and unexposed groups. Before matching, we expect to observe differences in the covariate means between the exposed and unexposed group because selection into exposures is non-random. We illustrated this problem for Study 4 in Table 5. We would like to know whether matching on the propensity score reduces the mean differences for covariates.

We normalize the difference of means for each variable using the pooled standard deviation of the covariate across exposed and unexposed groups. The normalized difference is preferred to a t-statistic that tests the null hypothesis of a difference in the means, because the t-statistic might be large if the sample is large, even if the difference is substantively small. Figure 6 presents the absolute standardized differences for each variable set in Study 4. In the upper left figure, only observables from variable set 1 (FB variables, see page 20 for a definition) are used to estimate the propensity scores to achieve balance. Although a number of the orange circles show a moderate difference before balancing (up to about 0.4 standard deviations), matching brings down the magnitude of these differences below 0.1 standard deviations. Stuart and Rubin (2007) suggest the standardized differences in means should not exceed 0.25. However, balancing on the FB variables does little to reduce the mean differences for the other variables, some of which have differences approaching one standard deviation.

In each of the next plots in Figure 6, we estimate the propensity score after including another set of variables. In each case, conditioning on this new set of variables successfully reduces the mean differences between the exposed and unexposed group for the added variables. The bottom right figure shows that conditioning on the full set of observables eliminates nearly all differences in means across variables. Thus, balancing on the propensity score also achieves balance on the covariate means across treatment groups for Study 4.

Detailed graphs of all remaining studies with the standardized differences obtained conditioning on all observables start on page A-4. Balancing on the propensity score substantially reduces the mean differences in all the studies and achieves good balance on the underlying observables.

Figure 6: Absolute standardized differences of covariate means for Study 4



[*] The sets of variables correspond to those described on page 20.

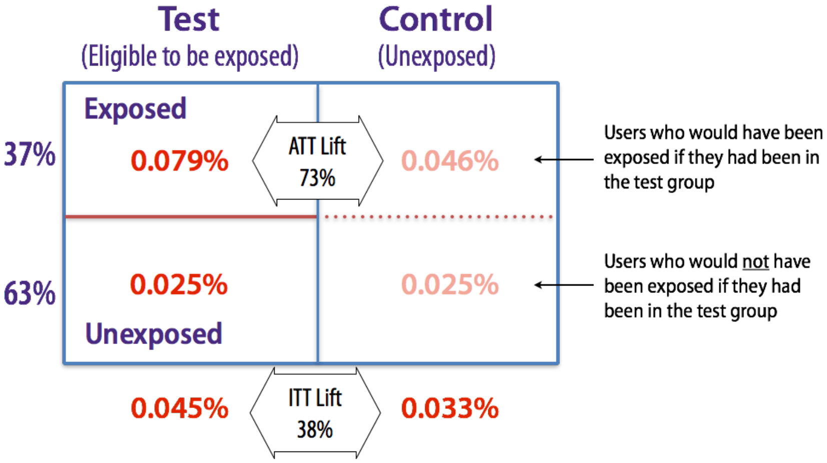
7 Results

We now present the results from the 15 studies. We first present the results from RCTs and then those from applying multiple observational approaches. Because of the data's confidential nature, all conversion rates have been scaled by a random constant, masking levels but allowing comparisons across studies.

7.1 RCT

To explain the results, we first highlight one typical advertising study (we refer to it as "Study 4"). An omnichannel retailer ran an ad campaign over two weeks in the first half of 2015. The study involved 25.5 million users randomly split into test and control groups in proportions of 70% and 30%, respectively. Ads were shown on mobile and desktop Facebook news feeds in the United States. For this study, the conversion pixel was embedded on the checkout-confirmation page. Therefore, the outcome measured in this study is whether a user purchased online during the study or up to several weeks after the study ended.[19]

Figure 7: Results from RCT


Test
(Eligible to be exposed)
Control
(Unexposed)
Exposed
37%
0.079%
ATT Lift
73%
0.046%
Users who would have been exposed if they had been in the test group
63%
0.025%
0.025%
Users who would not have been exposed if they had been in the test group
Unexposed
0.045%
ITT Lift
38%
0.033%

As Figure 7 shows, the conversion rates for the test and control groups of Study 4 were 0.045% and 0.033%, respectively, yielding an ITT of 0.012% and an ITT lift of 38%. We derive the estimated ATT by dividing the ITT (0.012%) by the percent of consumers who were exposed to the ad (37%), yielding an ATT of 0.033%. Based on the conversion rate of 0.079% for treated users in the test group, this implies the ATT lift was 73% (=0.033%/(0.079%-0.033%)). The 95% bootstrapped confidence interval for this lift is [49%, 103%].[20]

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

[20] Note that we do not know which users in the control group would have been exposed had they been assigned to the test group. This quantity is derived from the identifying assumption that had unexposed users in the test group been in the control groups, they would have had the same conversion probability (0.025%). Given that we know that

Table 3: ITT lift for all studies and measured outcomes

Study	Outcome	Conversion Prob. Control Group	Conversion Prob. Test Group	RCT ITT	RCT ITT Lift	RCT ITT Lift Confidence Interval
S1	Checkout	0.105%	0.131%	**0.027%**	**25.3%**	[13.8% 37.9%]
S2	Checkout	0.033%	0.033%	0.000%	1.0%	[-3.5% 5.7%]
S3	Checkout	0.203%	0.217%	**0.014%**	**6.9%**	[1.2% 13.0%]
S4	Checkout	0.033%	0.045%	**0.012%**	**37.7%**	[27.2% 49.1%]
S5	Checkout	0.009%	0.022%	**0.013%**	**153.2%**	[127.0% 182.4%]
S7	Checkout	0.247%	0.251%	0.004%	1.5%	[-0.2% 3.3%]
S8	Checkout	0.047%	0.047%	-0.001%	-1.2%	[-9.2% 7.6%]
S9	Checkout	0.184%	0.187%	**0.003%**	1.7%	[0.0% 3.5%]
S10	Checkout	0.113%	0.115%	0.002%	1.5%	[-9.1% 13.2%]
S11	Checkout	0.261%	0.277%	**0.016%**	**6.2%**	[3.3% 9.3%]
S12	Checkout	5.487%	5.547%	**0.059%**	**1.1%**	[0.1% 2.1%]
S13	Checkout	0.282%	0.272%	-0.010%	-3.5%	[-10.1% 3.5%]
S14	Checkout	0.027%	0.036%	**0.009%**	**34.3%**	[25.1% 44.1%]
S15	Checkout	1.385%	1.413%	**0.028%**	**2.0%**	[0.4% 3.7%]
S1	Registration	0.078%	0.570%	**0.492%**	**630.1%**	[568.3% 697.6%]
S5	Registration	0.078%	0.341%	**0.264%**	**339.9%**	[325.2% 355.2%]
S8	Registration	0.010%	0.012%	**0.003%**	**26.5%**	[5.8% 51.1%]
S10	Registration	0.363%	0.385%	0.022%	6.0%	[-0.2% 12.6%]
S14	Registration	0.165%	0.304%	**0.139%**	**84.5%**	[79.5% 89.7%]
S2	Page View	0.011%	0.123%	**0.111%**	**994.0%**	[917.7% 1076.1%]
S5	Page View	0.084%	0.275%	**0.191%**	**227.7%**	[216.8% 239.0%]
S6	Page View	0.356%	0.397%	**0.042%**	**11.7%**	[10.8% 12.5%]

RCT ITT and RCT ITT lift in **bold**: statistically different from zero at 5% level. 95% confidence intervals for RCT ITT Lift obtained via bootstrap.

Tables 3 and 4 present the results of the RCTs for all studies. The first table summarizes the ITT results; the second summarizes the ATT results. The results for Study 4, for example, are in the fourth row of each table.

Looking across all studies reveals a reasonable amount of variation in the percentage of the test group exposed to ads and in the ATT lift. Of the 14 studies with a checkout conversion, six failed to produce statistically significant lifts at the 5% significance level (although two were significant at a 10% level).

The lifts for registration and page-view outcomes are typically higher than for checkout outcomes,[21] for at least two possible reasons. One is that registration and page-view outcomes are easier outcomes to trigger via an advertisement, compared to a purchase—after all, the former outcomes typically require no payment. Second, specific registration and landing pages may be tied closely to the ad campaigns. Because unexposed users may not know how to get to the page, unexposed users are much less likely to reach that page than exposed users. For checkout outcomes, however, users in the control group can purchase from the advertiser as they normally would—triggering a conversion pixel does not take special knowledge of a page.[22]

the overall conversion probability of control users is 0.033% and that 37% of users were exposed, this implies the counterfactual conversion probability of exposed users in the test group is 0.046%.

[21] Johnson, Lewis, and Nubbemeyer (2017b) present a meta-study of 432 online display ad experiments on the Google Display Network. They find the median lift for site visits is 16% versus a median lift for purchases of 8%.

[22] One might ask why lifts for registration and page-view outcomes are not infinite, because—as we have just

Table 4: ATT lift for all studies and measured outcomes

		Pct	Conversion Prob.	Conversion Prob.	RCT ATT	RCT ATT	RCT ATT Lift	
Study	Outcome	Exposed	Exposed in Test	Unexposed in Test		Lift	Confidence	Interval
S1	Checkout	76%	0.151%	0.069%	**0.035%**	**30.0%**	[16%	46%]
S2	Checkout	48%	0.054%	0.014%	0.001%	1.3%	[-5%	8%]
S3	Checkout	66%	0.260%	0.131%	**0.021%**	**8.8%**	[1.1%	17%]
S4	Checkout	37%	0.079%	0.025%	**0.033%**	**72.8%**	[49%	103%]
S5	Checkout	30%	0.055%	0.008%	**0.045%**	**449.6%**	[306%	761%]
S7	Checkout	51%	0.284%	0.217%	0.007%	2.7%	[-0.3%	6%]
S8	Checkout	26%	0.069%	0.039%	-0.002%	-2.9%	[-21%	23%]
S9	Checkout	6.6%	2.105%	0.052%	**0.049%**	2.4%	[-0.1%	5%]
S10	Checkout	65%	0.127%	0.092%	0.003%	2.0%	[-11%	20%]
S11	Checkout	42%	0.488%	0.124%	**0.039%**	**8.6%**	[5%	13%]
S12	Checkout	77%	6.403%	2.810%	**0.078%**	**1.2%**	[0.2%	2%]
S13	Checkout	30%	0.187%	0.309%	-0.033%	-15.1%	[-35%	20%]
S14	Checkout	35%	0.068%	0.019%	**0.026%**	**62.0%**	[43%	86%]
S15	Checkout	81%	1.470%	1.175%	**0.034%**	**2.4%**	[0.4%	5%]
S1	Registration	76%	0.725%	0.064%	**0.643%**	**781.4%**	[694%	890%]
S5	Registration	30%	0.993%	0.068%	**0.893%**	**893.1%**	[797%	1010%]
S8	Registration	26%	0.025%	0.008%	**0.010%**	**63.2%**	[11%	176%]
S10	Registration	65%	0.423%	0.313%	0.033%	8.6%	[0%	19%]
S14	Registration	35%	0.642%	0.119%	**0.393%**	**158.1%**	[145%	173%]
S2	Page View	48%	0.249%	0.007%	**0.233%**	**1517.1%**	[1357%	1733%]
S5	Page View	30%	0.753%	0.075%	**0.647%**	**608.8%**	[541%	692%]
S6	Page View	61%	0.557%	0.152%	**0.069%**	**14.0%**	[13%	15%]

RCT ATT and RCT ATT Lift in **bold**: statistically different from zero at 5% level. 95% confidence intervals for RCT ATT Lift obtained via bootstrap.

Going forward, we will use the lift measured by the RCT as our gold standard for the truth, the benchmark against which to compare the observational methods.

7.2 Observational Models

Earlier we noted that we will evaluate observational methods by ignoring our experimental control group and analyzing only consumers in the test group. By doing so, we replicate the situation advertisers face when they rely on observational methods instead of an RCT, namely, to compare exposed to unexposed consumers, all of whom were in the ad's target group.

What selection bias do observational methods have to overcome to replicate the RCT results? Continuing with Study 4 as our example, Table 5 depicts the differences between unexposed and exposed users. For example, the second item there shows that exposed users are about eight percentage points more likely to be female than unexposed users. The table also demonstrates that, compared to unexposed users, exposed users are older, more likely to be married, have fewer Facebook friends, and tend to access Facebook more frequently from a mobile device than a desktop. As expected, a significant degree of covariate imbalance exists across the exposed and unexposed groups.

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 that shared the same landing or registration page.

Table 5: Mean by exposure

Variable	Unexposed	Exposed
Age	26.4	29.3
Female	88%	96%
Facebook Age	2202	2242
# of Friends	618	608
Facebook Web	5.1	4.0
Facebook Mobile	24.8	26.8
Married	8%	19%
Single	10%	14%
Phone A	5%	2%
Phone B	46%	52%
Phone C	48%	45%

If we are willing to (incorrectly) assume exposure is random, we could compare the exposed and unexposed groups, as in equation (13). The conversion rate among exposed and unexposed users was 0.061% and 0.019%, respectively, implying an ATT lift of 316%. This estimate represents the combined lift due to treatment *and* selection and is more than four times the lift due to treatment of 73%.

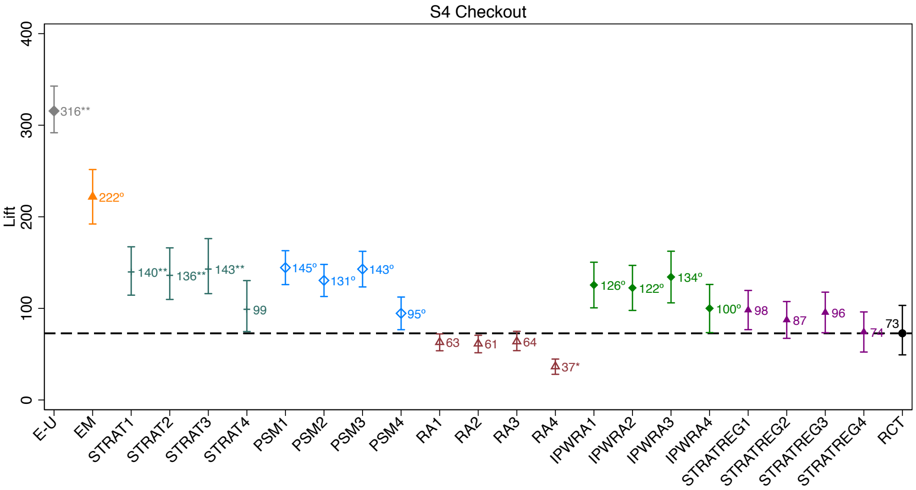
In the remainder of this section, we use multiple observational methods to estimate advertising effectiveness. Our goal is to assess how close these estimates come to the RCT ATT lift benchmark.

We first present three methods that match exposed and unexposed groups but do not rely on an outcome model: (1) exact matching (EM) based only on age and gender (as a naive benchmark used widely in industry), (2) stratification (STRAT), and (3) propensity score matching (PSM). The next three methods rely on an outcome model: (4) regression adjustment (RA) controls for observables but does not rely on matching, (5) inverse-probability-weighted regression adjustment (IPWRA) uses the propensity score to weigh observations in the regression model, and (6) stratification with regression (STRATREG) relies on both an outcome model and matching within strata. We begin with Study 4, and then highlight key findings with additional studies. Finally, we summarize the results of all 15 studies.

Study 4: Figure 8 summarizes the results from all methods. To interpret the graph, consider the right-most entry on the x-axis, "RCT." This graph shows the ATT lift from the RCT (73%) against which we compare all other methods. Next, the lift of each method is graphed with error bars. We describe the calculation of standard errors in section 6.2. We also report the results from a test of the hypothesis that the RCT lift equals the lift of a given method (see the graph for an explanation of symbols). In some cases, we cannot draw an inference if we are unable to compute the covariance between the RCT and observational treatment effects (e.g., using PSM).

Previously, we estimated a lift of 316% when comparing exposed with unexposed users (E-U). Given that this estimate represents the combined lift due to treatment and selection, it is not surprising that all methods come closer to the RCT. Exact matching (EM) on age and gender alone performs poorly, yielding a lift of 222%. All methods (except for the regression model, RA) show the following pattern: variable sets 1-3 overestimate lift compared to the RCT by about the

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



[**] and [*] means that we reject the hypothesis at a 1% or 5% significance level, respectively. [ ] means that we fail to reject the hypothesis. [^{O}] means that we cannot draw an inference.

same amount. The richest set of explanatory variables, variable set 4, yield estimates of lift that are close to the RCT, except for RA, which underestimates lift. Overall, whether a method relies on an outcome model seems unimportant.

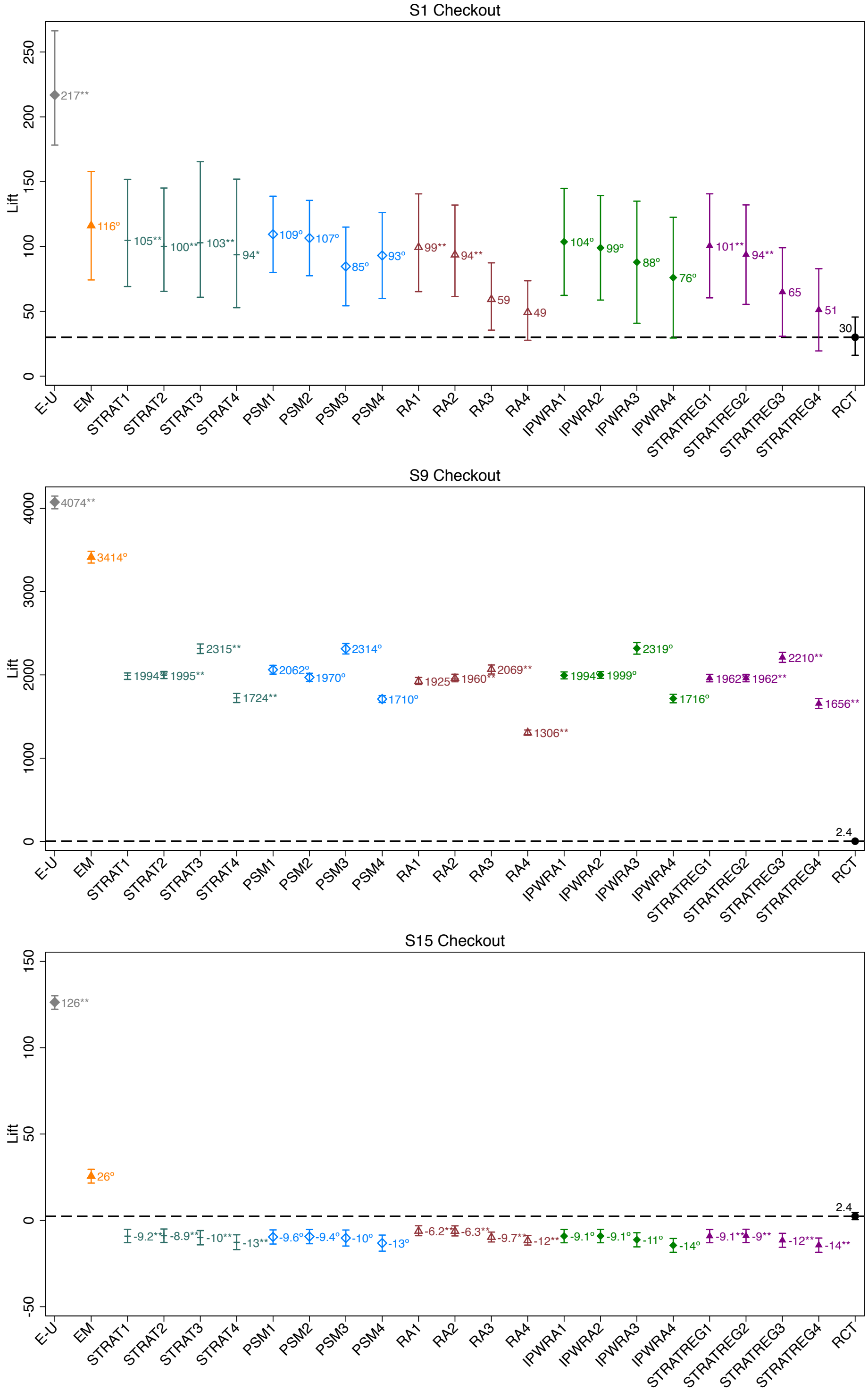
Study 4 suggests that some observational methods with a rich set of explanatory variables (e.g., STRATREG4) can recover the RCT lift. However, are the findings from Study 4 typical?

Study 1: Study 1's results follow a similar pattern to those in Study 4 (see top panel in Figure 9 on page 35). All methods do better than the exposed-unexposed comparison of 217%. As in Study 4, the best approaches (e.g., STRATREG4 or RA4) yield lift estimates statistically indistinguishable from the RCT.

Study 9: The pattern we observed in the two previous studies does not extend to other studies. Study 9, for example, has an RCT lift estimate of 2.4% (statistically different from 0 at a 10% level). The closest lift estimate is 1306% (RA4), a massive overestimate. We speculate that this discrepancy is partially the result of an unusually small exposure rate of only 6.6%, which leaves the ad-targeting mechanism ample opportunity to target the specific set of consumers most likely to respond. As in Studies 1 and 4, variable set 4 improves the lift estimate substantially, but it remains far from the RCT estimate.

Study 15: In Studies 1, 4, and 9, observational methods overestimated RCT lift for most methods and variable sets. However, as Study 15 shows, even this pattern is not generalizable. Except for exact matching on age and gender, all methods underestimate lift.

Figure 9: Summary of lift estimates and confidence intervals



S1 Checkout

S9 Checkout

S15 Checkout

[**] and [*] means we reject the hypothesis at a 1% or 5% significance level, respectively. [ ] means we fail to reject the hypothesis. [^{O}] means we cannot draw an inference.

Figure 10: Summary of lift results

		(A)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)
			EM	Propensity Score Matching				Regression Adjustment				Stratified Regression			
Campaign	Outcome	RCT Lift*	Age, Gender	Age, Gender + FB Vars	Age, Gender + FB Vars + Census Vars	Age, Gender + FB Vars + Census Vars + Activity Vars	Age, Gender + FB Vars + Census Vars + Activity Vars +FB Match Vars	Age, Gender + FB Vars	Age, Gender + FB Vars + Census Vars	Age, Gender + FB Vars + Census Vars + Activity Vars	Age, Gender + FB Vars + Census Vars + Activity Vars +FB Match Vars	Age, Gender + FB Vars	Age, Gender + FB Vars + Census Vars	Age, Gender + FB Vars + Census Vars + Activity Vars	Age, Gender + FB Vars + Census Vars + Activity Vars +FB Match Vars
1	Checkout	30%	116%	109%	107%	85%	93%	104%	99%	88%	76%	101%	94%	65%	51%
2	Checkout	1.3%	432%	161%	149%	37%	36%	149%	140%	43%	35%	97%	98%	54%	40%
3	Checkout	8.8%	65%	20%	24%	41%	17%	21%	23%	38%	5%	18%	19%	30%	2%
4	Checkout	73%	222%	145%	131%	143%	95%	126%	122%	134%	100%	98%	87%	96%	74%
5	Checkout	450%	511%	418%	443%	463%	316%	428%	432%	437%	305%	447%	431%	435%	301%
7	Checkout	2.7%	37%	20%	18%	-33%	-36%	19%	20%	-33%	-35%	19%	19%	-31%	-33%
8	Checkout	-2.9%	48%	31%	36%	50%	27%	36%	41%	54%	29%	32%	37%	52%	28%
9	Checkout	2.4%	3414%	2062%	1970%	2314%	1710%	1994%	1999%	2319%	1716%	1962%	1962%	2210%	1656%
10	Checkout	2.0%	38%	23%	16%	43%	-7%	20%	20%	34%	-13%	21%	21%	35%	-11%
11	Checkout	9%	275%	29%	31%	38%	7%	30%	31%	35%	3%	30%	31%	34%	2%
12	Checkout	1%	129%	111%	110%	82%	82%	112%	111%	82%	81%	112%	111%	84%	82%
13	Checkout	-15%	-39%	-35%	-36%	-30%	-31%	-35%	-35%	-31%	-30%	-35%	-35%	-31%	-30%
14	Checkout	62%	119%	80%	85%	95%	101%	80%	83%	92%	90%	74%	77%	82%	84%
15	Checkout	2%	26%	-10%	-9%	-10%	-13%	-9%	-9%	-11%	-14%	-9%	-9%	-12%	-14%
1	Registration	781%	1024%	978%	944%	1060%	977%	968%	960%	1087%	985%	824%	800%	432%	348%
5	Registration	893%	1270%	1071%	1055%	1070%	765%	1067%	1067%	1063%	728%	1112%	1104%	1081%	772%
8	Registration	63%	180%	162%	159%	173%	167%	150%	153%	158%	114%	157%	161%	160%	125%
10	Registration	9%	34%	19%	18%	34%	-3%	18%	18%	31%	0%	19%	18%	31%	2%
14	Registration	158.1%	275%	215%	219%	244%	241%	219%	219%	238%	234%	219%	218%	240%	239%
2	Page View	1517%	4261%	2493%	2416%	1150%	1177%	2408%	2422%	1175%	1187%	1162%	1181%	1722%	1268%
5	Page View	609%	846%	771%	731%	719%	484%	751%	748%	710%	477%	776%	769%	717%	498%
6	Page View	14%	227%	103%	105%	263%	255%	103%	106%	250%	246%	111%	115%	255%	278%

* Red: RCT Lift is statistically different from 0 at 5% significance level

🟥 Observational method overestimates lift

🟦 Observational method underestimates lift

Color proportional to overestimation factor; darkest color reached at 3-times over- or underestimation

7.3 Summary of All 15 Studies

Detailed graphs with the results of all remaining studies start on page A-1. A summary of nearly all results in these graphs is in Figure 10. The rows correspond to a study-conversion-type pair. The first results column reports the RCT lift, which is highlighted in red if it is statistically significant at a 5% level. Remaining columns contain the lift estimates of the observational methods we analyze. The sequence of columns corresponds to the sequence of methods in the detailed graphs. Each cell is color coded to represent when and by how much observational lift estimates differ from RCT lift estimates. Red (blue) means the observational method overestimates (underestimates) lift. The darkest shade means the observational method over- or underestimates the RCT lift by a factor of 3 or more. The color is proportional to the magnitude of misestimation.

A scan of Figure 10 reveals several clear patterns. First, the observational methods we study mostly overestimate the RCT lift, although in some cases, they can significantly underestimate RCT lift. Second, the point estimates in seven of the 14 studies with a checkout-conversion outcome are consistently off by more than a factor of three. Third, observational methods do a better job of approximating RCT outcomes for registration and page-view outcomes than for checkouts. We believe the reason is the nature of these outcomes. Because unexposed users (both treatment and control) are relatively unlikely to find a registration or landing page on their own, comparing the exposed group in treatment with 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 with the (always unexposed) control group (the comparison the RCT is based on). Fourth, scanning across approaches, because the exposed-unexposed comparison represents the combined treatment and selection effect—given the nature of selection in this industry—the estimate is always strongly biased up, relative to the RCT lift. Exact matching on gender and age decreases that bias, but it remains significant. Generally, we find that more information helps, but adding census data and activity variables helps less than the Facebook match variable. We do not find that one method consistently dominates: in some cases, a given approach performs better than another for one study but not the other.

8 Assessing the Role of Unobservables in Reducing Bias

Many of the observational models are unable to recover the RCT treatment effect, even though matching on the propensity score achieves good balance on the propensities themselves and on the underlying distribution of covariates (see section 6.3). Of course, the unconfoundedness assumption requires not only balance on observables, but also that no unobservables exist that might be correlated with treatment and potential outcomes.

Next, we present a sensitivity analysis based on the following thought experiment: "If we could obtain new observables, how much better would they need to be to eliminate the bias between the observational and RCT estimates?" This analysis is motivated by the fact that, although our data are rich by certain standards, we do not observe all the information Facebook uses to run its advertising platform. Hence, this section investigates whether additional data might possibly eliminate the bias between the observational and RCT estimates.

To help frame our approach, suppose unconfoundedness fails, such that

$$(Y_i(0), Y_i(1)) \not\perp\!\!\!\perp W_i | X_i \ .$$

Now suppose some unobservable $U \in \mathbb{R}$ exists correlated with Y and W such that—if we observed U—unconfoundedness would once again hold,

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i | X_i, U_i \ . \tag{30}$$

The magnitude of the bias from ignoring U depends on the strength of its correlations with Y and W. Our approach, based on the methodology developed in Rosenbaum and Rubin (1983a) and extended by Ichino, Fabrizia, and Nannicini (2008) (see page A-14 for details),[23] is to simulate an unobservable U that eliminates this bias.

Once we have simulated the unobservable U for a study, we compare the explanatory power of U with the explanatory power of our observables to assess the likelihood that such data could actually be obtained. Let $R^2_Y(X, U)$ be the R^2 from a regression of the outcome on (X, U) for the subset of observations with $W_i = 0$. Similarly, $R^2_Y(X) \equiv R^2_Y(X, 0)$ is the R^2 from a regression on X alone for untreated users. The relative strength of the unobservable to affect outcomes is

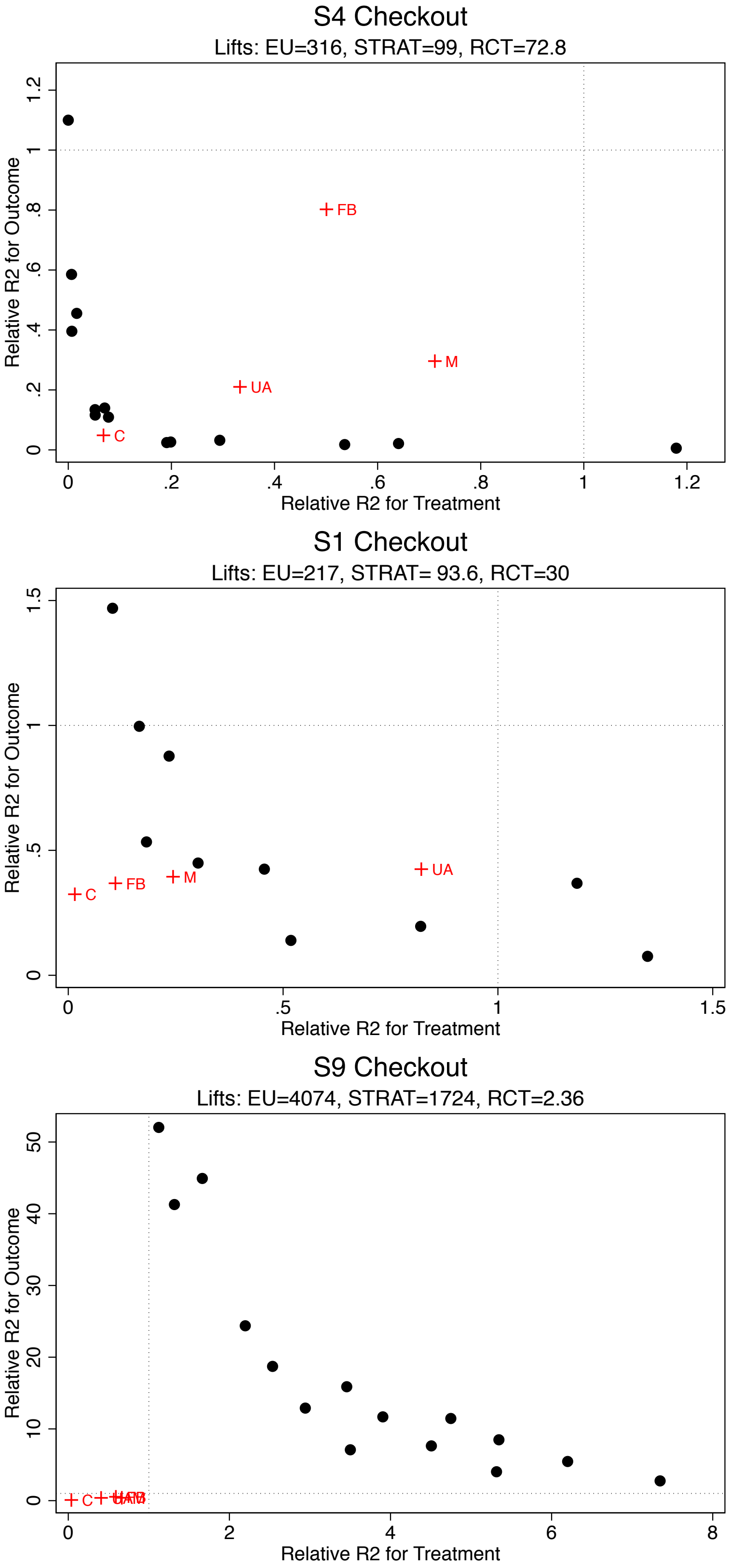
$$R^2_{Y,rel} = \frac{R^2_Y(X, U) - R^2_Y(X)}{R^2_Y(X)}$$

and similar for $R^2_{W,rel}$. The interpretation is as follows: $R^2_{Y,rel} = 1$, and $R^2_{W,rel} = 1$ represents the combined power of all four variable sets in explaining variation in outcome and treatment, respectively. Hence, if we found an unobservable with, for example, $R^2_{Y,rel} = 2$, and $R^2_{W,rel} = 3$, this unobservable would have to explain two times as much of the outcome variation and three times as much of the treatment variation as our combined variable sets to eliminate the bias in the observational method.

The top panel in Figure 11 presents the sensitivity analysis for Study 4. As the subtitle of the chart suggests, estimating using stratification produces a treatment with a lift estimate of 99% versus the RCT estimate of 73%. Notice the combined treatment and selection effect (measured through the exposed-unexposed comparison) is 316%, which means our observational data have succeeded in eliminating much but not all of the selection effect. The horizontal axis characterizes the relative R^2 for treatment and the vertical axis characterizes the relative R^2 for outcomes. The strength of our observables is displayed using "+", obtained from separate regressions of the particular variable set on treatment and outcomes. For example, the user activity (UA) variables alone explain about 35% of the relative variation for treatment and 20% of the relative variation for outcomes. The black dots represent points at which using the unobservable was able to generate

[23] A related literature assesses the strength of selection effects in observational settings, usually assuming selection on X informs the degree of selection on U. See Murphy and Topel (1990), Altonji, Elder, and Taber (2005), and Oster (Forthcoming).

Figure 11: Sensitivity Analysis for Studies 4, 1, and 9



[*] "FB" are the FB variables, "C" are the Census variables, "UA" are the user-activity variables, "M" are the match-score variables.

the RCT treatment effect. Not surprisingly, given that the degree of remaining bias after using stratification is small relative to the bias reduction that stratification already achieved with our observables, a relatively weak unobservable is required to remove the bias entirely. The unobservable could be the same strength as the census data, which explains about 7% of the relative treatment variation and 5% of the relative outcome variation. Alternatively, the unobservable could explain less of the variation in outcome if it increases its explained variation in treatment.

The sensitivity analysis for Study 1 in Figure 11 presents a somewhat different picture. In this study, using stratification produces a treatment with a lift estimate of 93.6% versus the RCT estimate of 30%. The combined treatment and selection effect is 217%, which means our observational data have eliminated less of the selection effect than in Study 4. Here a stronger unobservable is required to remove the bias entirely. The unobservable could be the same strength as variable set 4 (M in the graph), which explains about 40% of the relative treatment and outcome variation.

As our final example, consider a study in which stratification leads to massively biased estimates of the RCT lift, 1724% relative to the 2.36%, with a combined treatment and selection effect of 4047%. As the bottom right panel of Figure 11 shows, the unobservable would need to have between 5 and 10 times as much explanatory power as the observables in our data.

Detailed graphs with the results of all remaining studies start on page A-17. We analyze only checkout-conversion outcomes, because they are the outcomes for which observational methods performed the worst. Inspecting the results yields a number of additional insights. Let *brr* (bias reduction ratio) be the ratio of remaining bias after stratification and the total selection effect:

$$brr = \frac{|\text{STRAT Lift -RCT Lift}|}{|\text{EU Lift-RCT Lift}|} .$$

Table 6 sorts the studies by this ratio. Visual inspection of the graphs on page A-17 shows that the degree of additional information needed to eliminate the bias increases roughly with *brr*. Studies that need little additional information are 11, 3, 2, 4, 14, and 10. Studies in which one would need

Table 6: Summary of bias reduction through stratification

Study	Conversion Outcome	EU Lift	STRAT Lift	RCT Lift	brr
11	checkout	392%	7.1%	8.6%	0.4%
3	checkout	198%	18%	8.8%	5%
2	checkout	377%	37%	1.3%	10%
4	checkout	316%	99%	73%	11%
14	checkout	365%	99%	62%	12%
15	checkout	126%	-13%	2.4%	12%
10	checkout	138%	-15%	2%	13%
8	checkout	179%	33%	-2.9%	20%
13	checkout	61%	-30%	-15%	20%
7	checkout	131%	35%	2.7%	25%
1	checkout	217%	94%	30%	34%
12	checkout	233%	81%	1.2%	34%
9	checkout	4074%	1724%	2.4%	42%
5	checkout	678%	306%	450%	63%

massive additional information relative to what we observe are 12 and 9. Studies 15, 8, 13, 7, 1, and 5 fall roughly in the middle; they require data on the order of one or two of the variable sets we have.

We also gain additional insights about the nature of our explanatory variables. First, the census variables (C) generally explain little of the variation; moreover, they never account for more than 10% of explained variation in treatment exposure, but in 6 of 14 cases account for between 10% and 60% of the explained variation in the outcome. Second, the user-activity (UA) variable mostly explains treatment exposure rather than outcome. This is consistent with the finding in Lewis, Rao, and Reiley (2011) that unobserved user activity led to selection into exposure. We also find that the match (M) variable mostly explains treatment exposure rather than outcome, which is not ex-ante obvious, given that this variable is a composite of many user characteristics and behaviors. Third, the Facebook (FB) variables generally have high explanatory power that applies similarly to treatment exposure and outcome.

Our results show that for some studies, observational methods would require additional covariates that exceed considerably our combined observables' explanatory power. This suggests that eliminating bias from observational methods would be hard, even for industry insiders with access to additional data.

9 Conclusion

In this paper, we have analyzed whether the variation in data typically available in the advertising industry enables observational methods to substitute reliably for randomized experiments in online advertising measurement. We have done so by using a collection of 15 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 with the results obtained from the RCT.

We find that across the advertising studies, on average, a significant discrepancy exists between the observational approaches and RCTs. The observational methods we analyze mostly overestimate the RCT lift, although in some cases, they significantly underestimate this lift. The bias can be high: in 50% of our studies, the estimated percentage increase in purchase outcomes is off by a factor of three across all methods. With our small number of studies, we could not identify campaign characteristics that are associated with strong biases. We also find that observational methods do a better job of approximating RCT lift for registration and page-view outcomes than for purchases. Finally, we do not find that one method consistently dominates. Instead, a given approach may perform better for one study but not another.

Our paper makes three contributions. The first is to shed light on whether—as is thought in the industry—sophisticated observational methods based on the individual-level data plausibly attainable in the industry are good enough for ad measurement, or whether these methods likely yield unreliable estimates of the causal effects of advertising. Results from our 15 studies support the latter: the methods we study yield biased estimates of causal effects of advertising in a majority of cases. In contrast to existing examples in the academic literature, we find evidence of both under- and overestimates of ad effectiveness. These biases persist even after conditioning on a rich set of observables and using a variety of flexible estimation methods.

Our second contribution is to characterize the nature of the unobservable needed to use observational methods successfully to estimate ad effectiveness. Specifically, we conduct a thought experiment to characterize the quality of data required, above and beyond our current data, to allow an observational model to recover the RCT treatment effect. In more than half the cases, the additional data would need to be as strong as one or two of our better-performing variables sets; obtaining such data is likely not trivial.

Third, we add to the literature on observational versus experimental approaches to causal measurement. Over the last two decades, we have seen significant improvements in observational methods for causal inference (Imbens and Rubin 2015). We 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. We found they do not—at least not with the data at our disposal.

One caveat related to our conclusion is that the performance of the observational methods we study is only as good as the data. Our data possess a number of strengths relative to other online advertising studies: the sheer size of each ad experiment, the rich set of observables, and Facebook's ability to track exposures and conversions across all of a user's devices. Moreover, Facebook's closed system makes solving selection issues an easier problem than many other ad-effectiveness applications because advertisers more likely base their bids on the same information sets. By contrast, in display ads purchased via real-time bidding (RTB), advertisers often have private information, from their own businesses or through third-party data providers, to inform their bidding strategies. Nonetheless, our data do not encompass all the data Facebook relies on in its ad-delivery system, nor does Facebook necessarily log or retain all these data for future analysis. Better data, potentially requiring additional logging, could help improve the performance of observational methods.

References

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.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113(1), 151–184.

Arceneaux, K., A. S. Gerber, and D. P. Green (2010): "A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark," *Sociological Methods and Research*, 39(2), 256–282.

Athey, S., G. W. Imbens, and S. Wager (forthcoming): "Approximate Residual Balancing: De-biased Inference of Average Treatment Effects in High Dimensions," *Journal of the Royal Statistical Society: Series B*.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80(6), 2369–2429.

Benady, D. (2016): "Can Advertising Support A Free Internet?," *The Guardian*.

Berkovich, P., and L. Wood (2016): "Using Single-Source Data to Measure Advertising Effectiveness," Discussion Paper 2, Nielsen, http://www.nielsen.com/us/en/insights/reports/2016/using-single-source-data-to-measure-advertising-effectiveness.html.

Bickel, P., F. Gotze, and W. van Zweet (1997): "Resampling Fewer than n observations: Gains, Losses, and Remedies for Losses," *Statistica Sinicia*, 7, 1–31.

Blake, T., C. Nosko, and S. Tadelis (2015): "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment," *Econometrica*, 83(1), 155–174.

Bond, D. (2017): "Advertising agencies squeezed by tech giants," *The Financial Times*.

Bronnenberg, B. J., J.-P. Dubé, and C. F. Mela (2010): "Do Digital Video Recorders Influence Sales?," *Journal of Marketing Research*, 47(6), 998–1010.

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

Cochran, W. G. (1968): "The Effectiveness of Adjustment by Subclassifcation in Removing Bias in Observational Studies," *Biometrics*, 24(2), 295–314.

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

Cooper, M. J., H. Gulen, and P. R. Rau (2005): "Changing Names with Style: Mutual Fund Name Changes and Their Effects on Fund Flows," *The Journal of Finance*, 60(6), 2825–2858.

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

Draganska, M., W. R. Hartmann, and G. Stanglein (2014): "Internet Versus Television Advertising: A Brand-Building Comparison," *Journal of Marketing Research*, 51, 578–590.

Eckles, D., and E. Bakshy (2017): "Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects," *Working paper, MIT Sloan School of Management.*

Ellickson, P., R. L. Collins, K. Hambarsoomians, and D. F. McCaffrey (2005): "Does alcohol advertising promote adolescent drinking? Results from a longitudinal assessment," *Addiction*, 100(2), 235–246.

Ferraro, P. J., and J. J. Miranda (2014): "The performance of non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark," *Journal of Economic Behavior & Organization*, 107, 344–365.

——— (2017): "Panel Data Designs and Estimators as Substitutes for Randomized Controlled Trials in the Evaluation of Public Programs," *Journal of the Association of Environmental and Resource Economists*, 4(1), 281–317.

Gelman, A., and J. Hill (2007): *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 1st edn.

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

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

Gu, X., and P. Rosenbaum (1993): "Journal of Computational and Graphical Statistics," *Comparison of multivariate matching methods: structures, distances, and algorithms*, 2, 405–420.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Ichino, A., M. Fabrizia, and T. Nannicini (2008): "From Temporary Help Jobs to Permanent Employment: What can we Learn from Matching Estimators and Their Sensitivity?," *Journal of Applied Econometrics*, 23, 305–327.

Imai, K., and M. Ratkovic (2014): "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society, Series B*, 76(1), 243–263.

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

IMBENS, G. W. (2003): "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93(2), 126–132.

——— (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., 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. A. LEWIS, AND E. I. NUBBEMEYER (2017a): "Ghost Ads: Improving the Economics of Measuring Ad Effectiveness," *Journal of Marketing Research*, 54(6), 867–884.

——— (2017b): "The Online Display Ad Effectiveness Funnel & Carryover: Lessons from 432 Field Experiments," Working paper.

JOHNSON, G. A., R. A. LEWIS, AND D. REILEY (2016): "Location, Location, Location: Repetition and Proximity Increase Advertising Effectiveness," Working paper, University of Rochester.

——— (2017): "When Less is More: Data and Power in Advertising Experiments," *Marketing Science*, 36(1), 43–53.

KALYANAM, K., J. MCATEER, J. MAREK, J. HODGES, AND L. LIN (2018): "Cross Channel Effects of Search Engine Advertising on Brick & Mortar Retail Sales: Meta Analysis of Large Scale Field Experiments on Google.com," *Quantitative Marketing and Economics*, 16(1), 1–42.

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

KUMAR, A., R. BEZAWADA, R. RISHIKA, R. JANAKIRAMAN, AND P. KANNAN (2016): "From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior," *Journal of Marketing*, 80(1), 7–25.

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 J. RAO (2015): "The unfavorable economics of measuring the returns to advertising," *Quarterly Journal of Economics*, 130(4), 1941–1973.

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.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004): "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods*, 9(4), 403–425.

Murphy, K., and R. Topel (1990): "Eciency Wages Reconsidered: Theory and Evidence," in *Advances in the Theory and Measurement of Unemployment*, ed. by Y. Weiss, and G. Fishelson, vol. 2, pp. 204–240. Palgrave Macmillan UK.

Oster, E. (Forthcoming): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business and Economic Statistics*.

Pham, T. T., and Y. Shen (2017): "A Deep Causal Inference Approach to Measuring the Effects of Forming Group Loans in Online Non-profit Microfinance Platform," Working paper, Stanford GSB, available at https://arxiv.org/pdf/1706.02795.pdf.

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. B. Rubin (1983a): "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society, Series B*, 45(2), 212–218.

——— (1983b): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrica*, 70, 41–55.

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

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

——— (2001): "Using propensity scores to help design observational studies: application to the tobacco litigation," *Health Services and Outcomes Research Methodology*, 2, 169–188.

Rubin, D. B., and R. P. Waterman (2007): "Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology," *Statistical Science*, 21(2), 206–222.

Sahni, N. (2015): "Effect of Temporal Spacing between Advertising Exposures: Evidence from Online Field Experiments," *Quantitative Marketing and Economics*, 13(3), 203–247.

Sahni, N., and H. Nair (2016): "Native Advertising, Sponsorship Disclosure and Consumer Deception: Evidence from Mobile Search-Ad Experiments," Working paper, Stanford GSB, available at SSRN: http://ssrn.com/abstract=2737035.

Stuart, E. A. (2010): "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, 25(1), 1–21.

Stuart, E. A., and D. B. Rubin (2007): *Matching methods for causal inference: Designing observational studies*chap. Best Practices in Quantitative Methods. Sage Publishing (Thousand Oaks, CA).

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, 58(1), 267–288.

Westreich, D., J. Lessler, and M. J. Funk (2010): "Propensity score estimation: machine learning and classification methods as alternatives to logistic regression," *Journal of Clinical Epidemiology*, 63(8), 826–833.

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

Zubizarreta, J. R. (2012): "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery," *Journal of the American Statistical Association*, 107, 1360–1371.

——— (2015): "Stable weights that balance covariates for estimation with incomplete outcome data," *Journal of the American Statistical Association*, 110(511), 910–922.

ONLINE APPENDIX

Additional figures

Figure A-1: Checkout conversions: Estimated propensity scores by treatment status and pre/post matching studies 1-5, 7

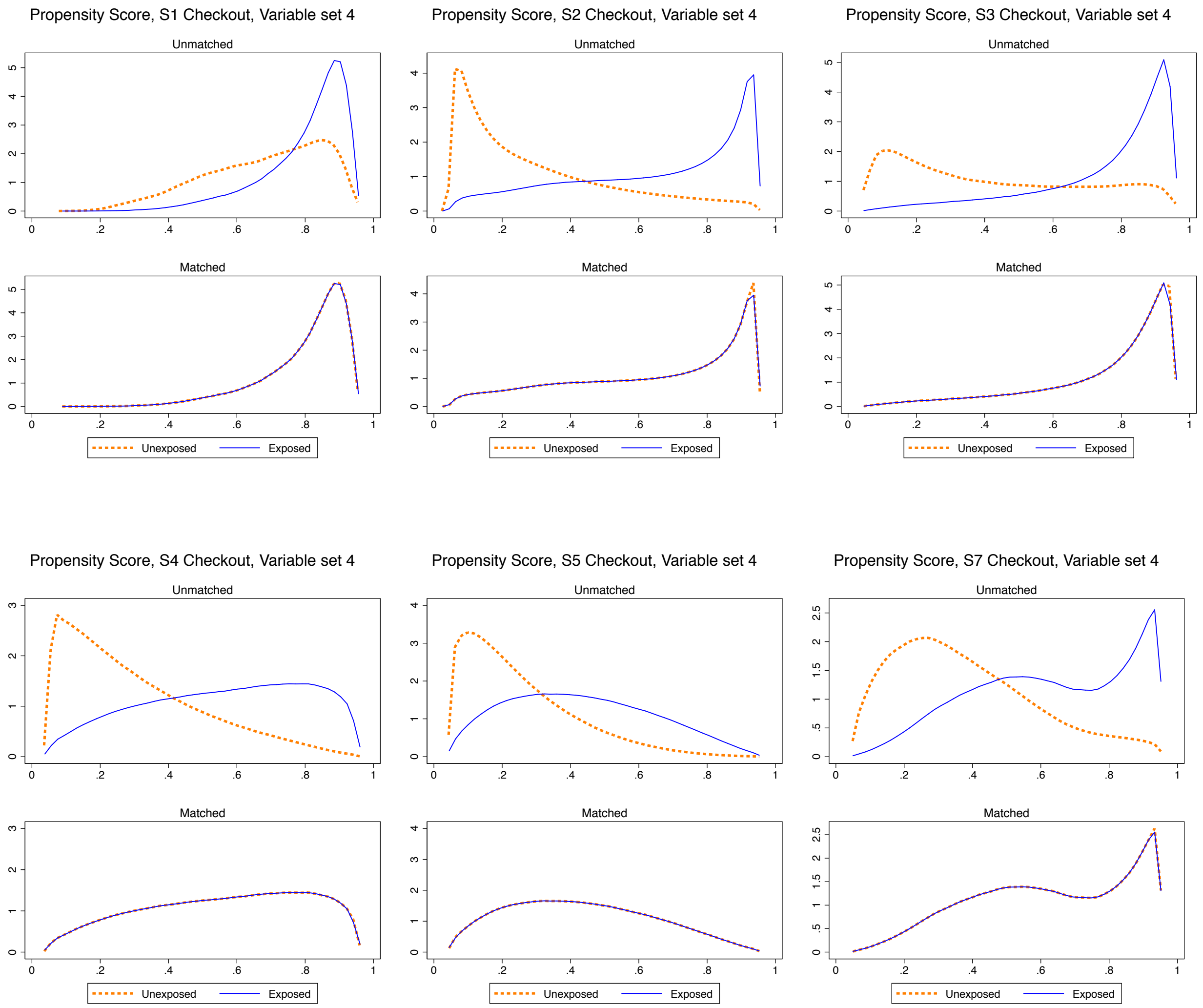


Figure A-2: Checkout conversions: Estimated propensity scores by treatment status and pre/post matching studies 8-15

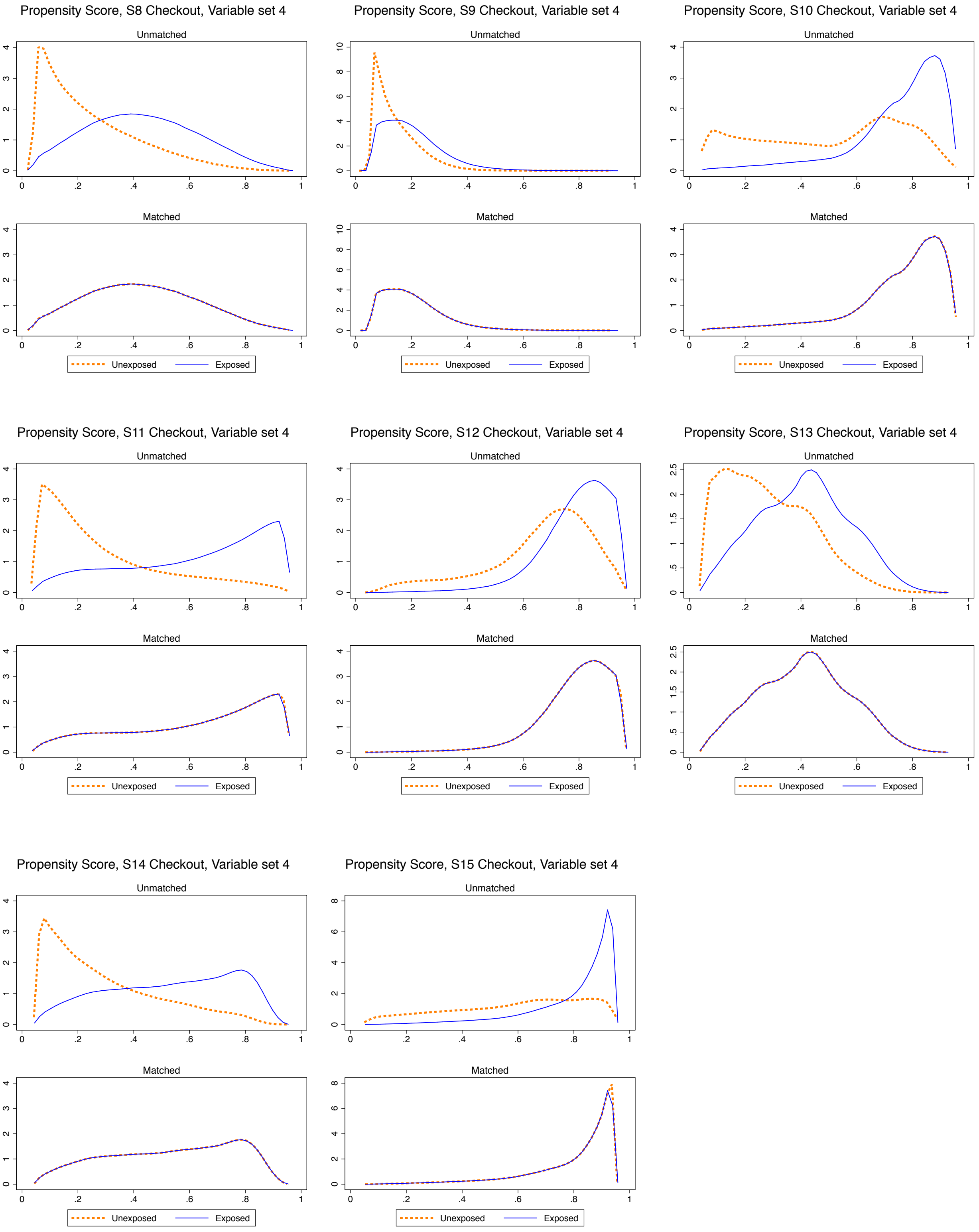


Figure A-3: Registration conversions, Studies 1, 5, 8, 10, 14; and page-view conversions, Studies 2, 5, 6: Estimated propensity scores by treatment status and pre/post matching

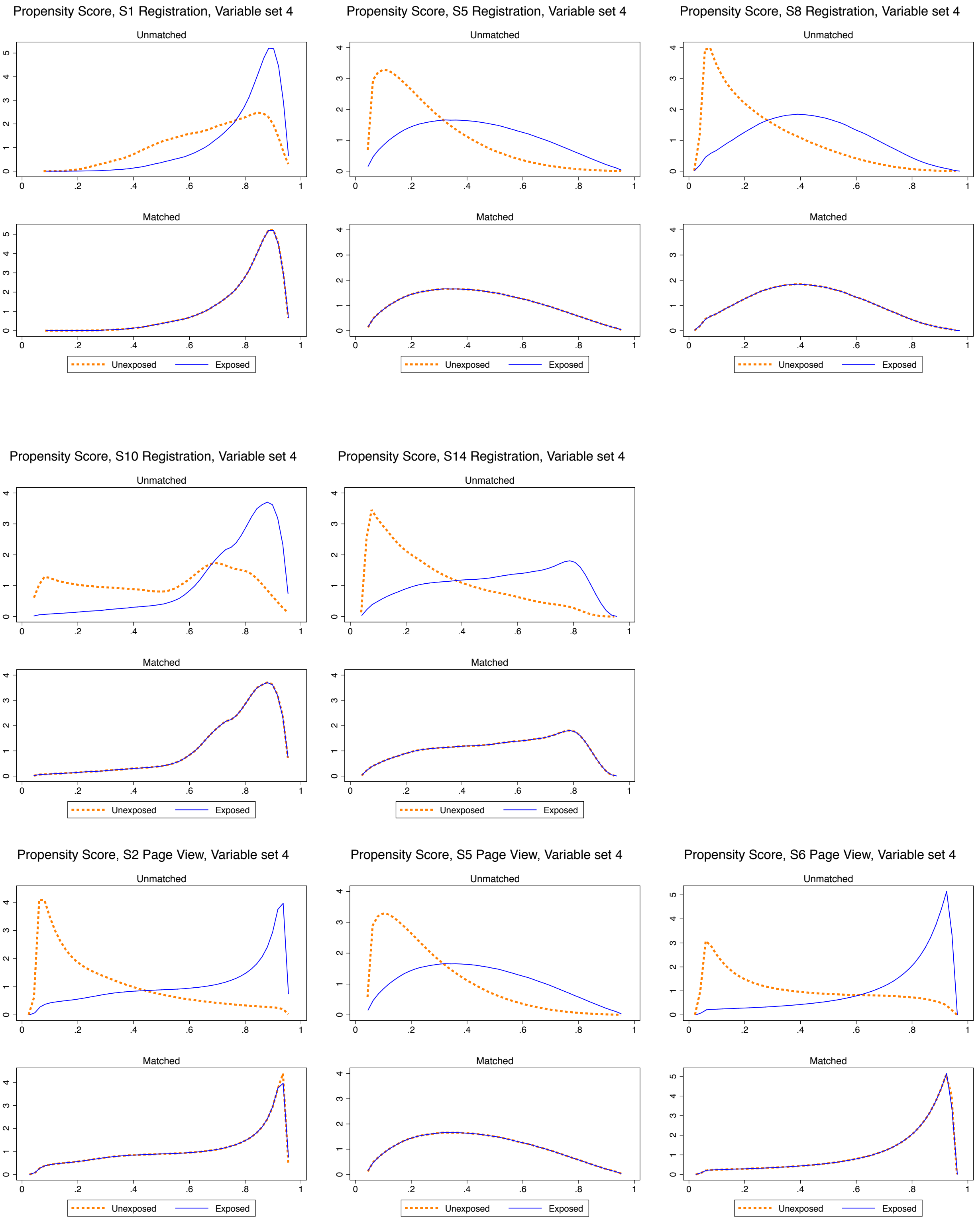


Figure A-4: Checkout conversions: Absolute standardized differences of covariate means for Studies 1-5, 7-9

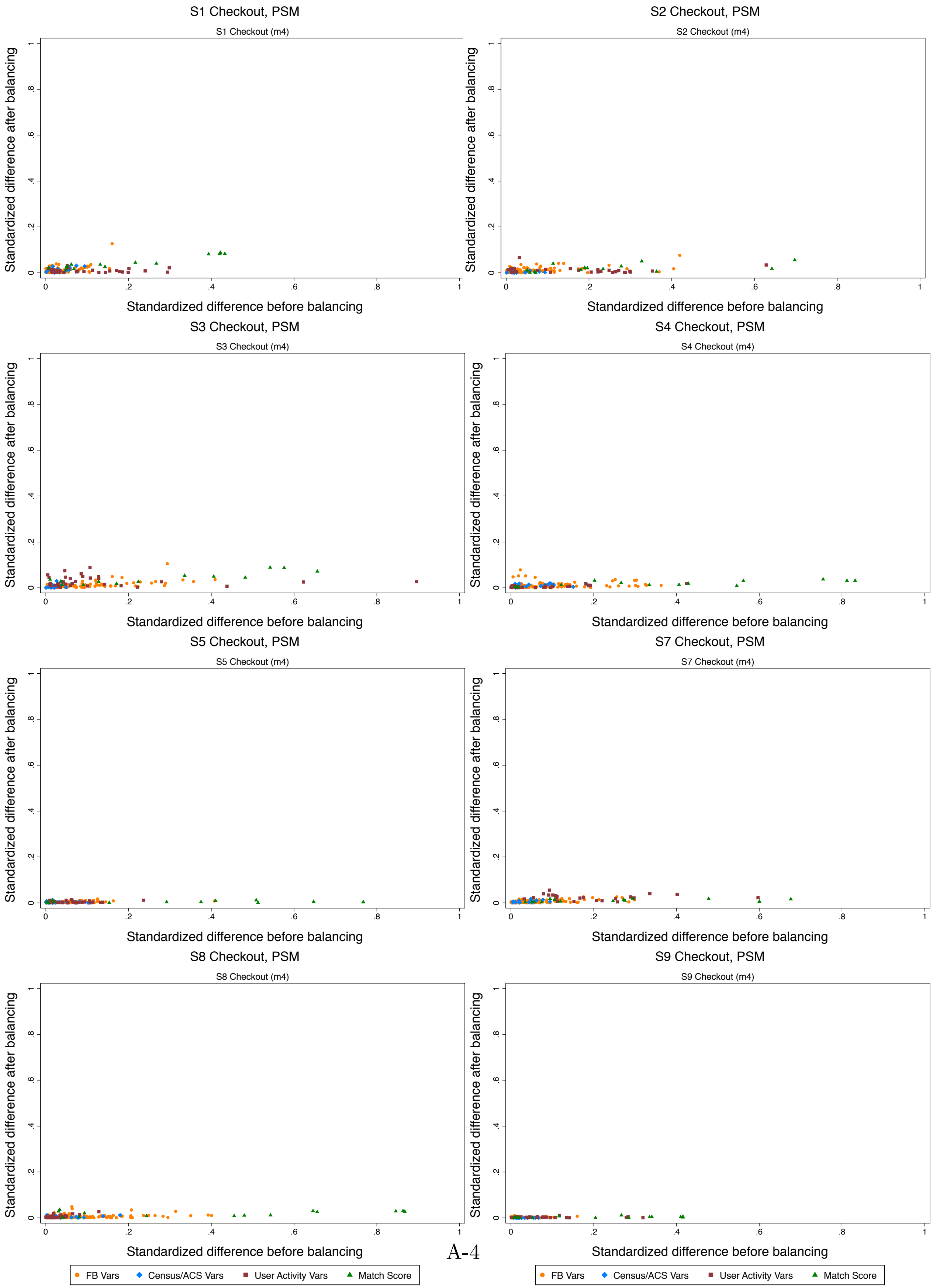


Figure A-5: Checkout conversions: Absolute standardized differences of covariate means for Studies 10-15

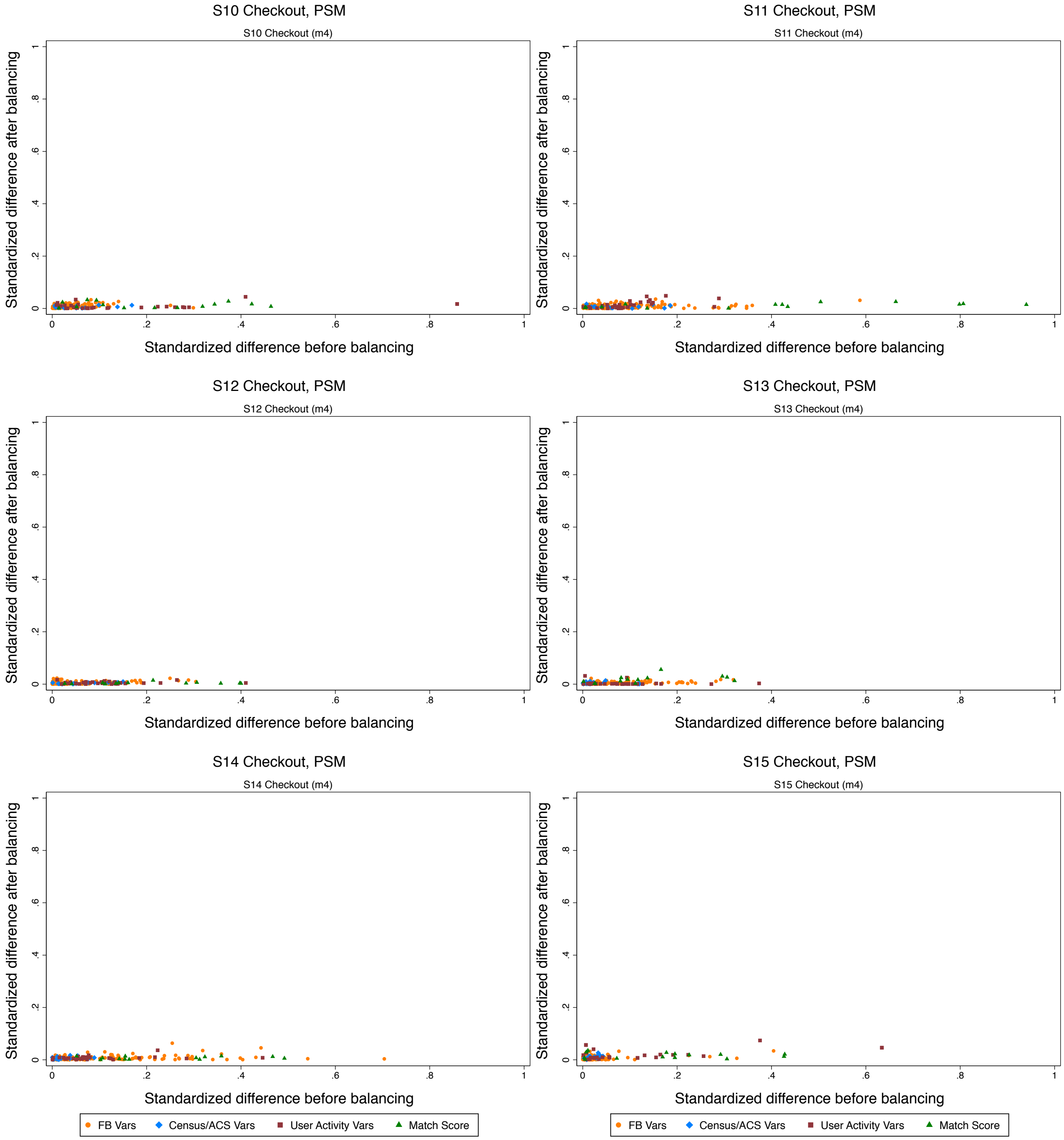


Figure A-6: Registration conversions: Absolute standardized differences of covariate means for Studies 1, 5, 8, 10, and 14

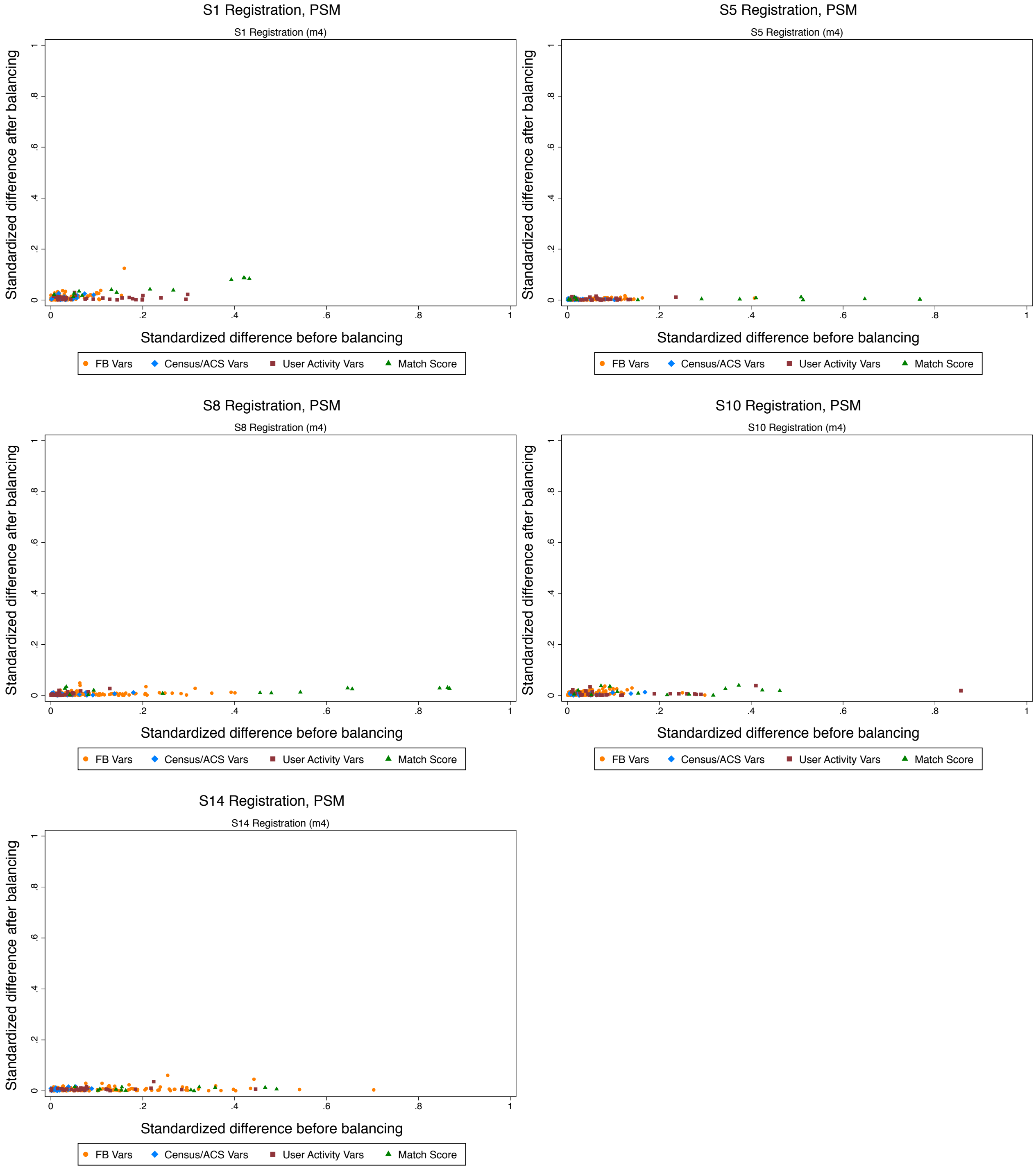


Figure A-7: Page-view conversions: Absolute standardized differences of covariate means for Studies 2, 5, and 6

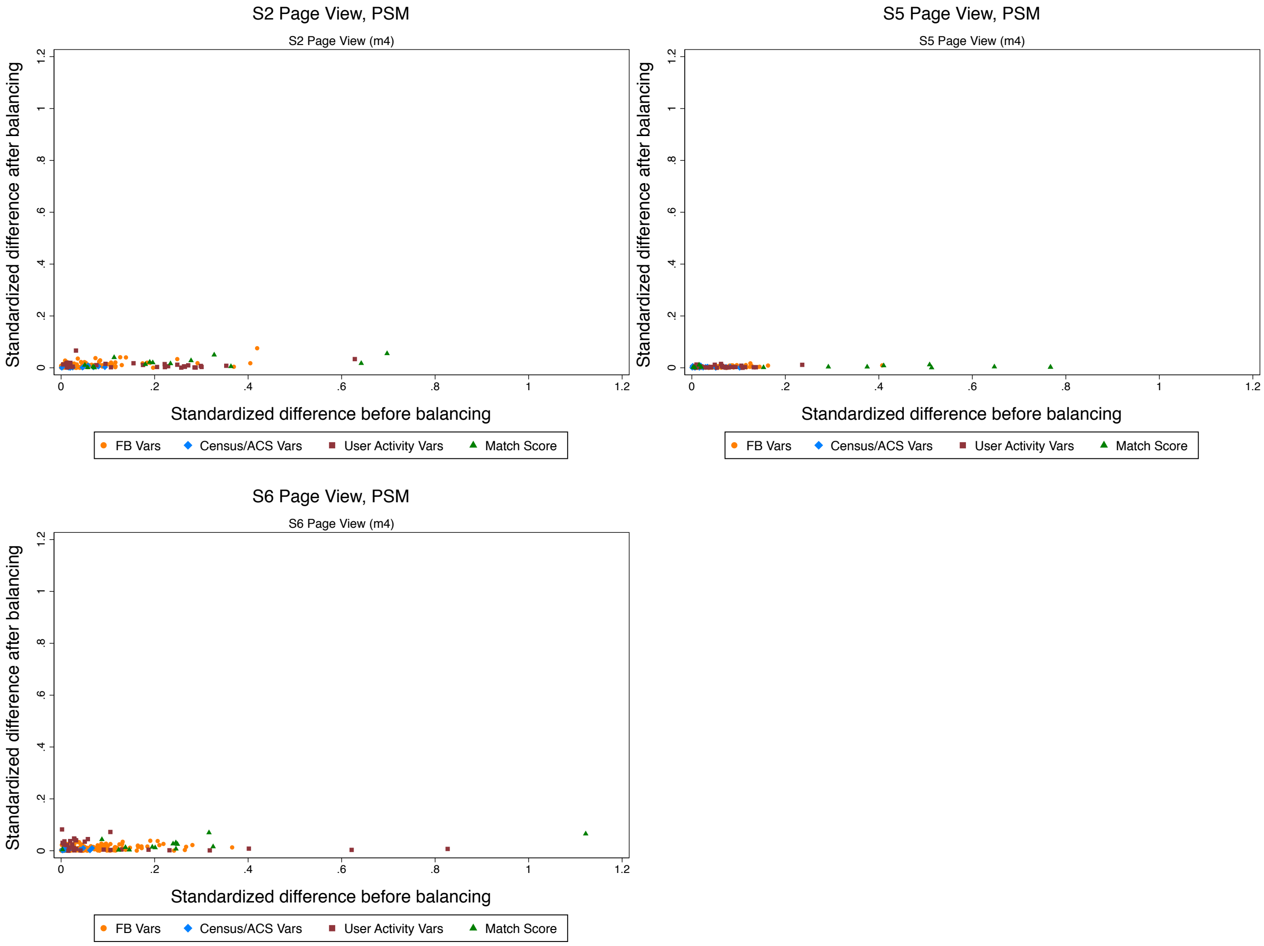
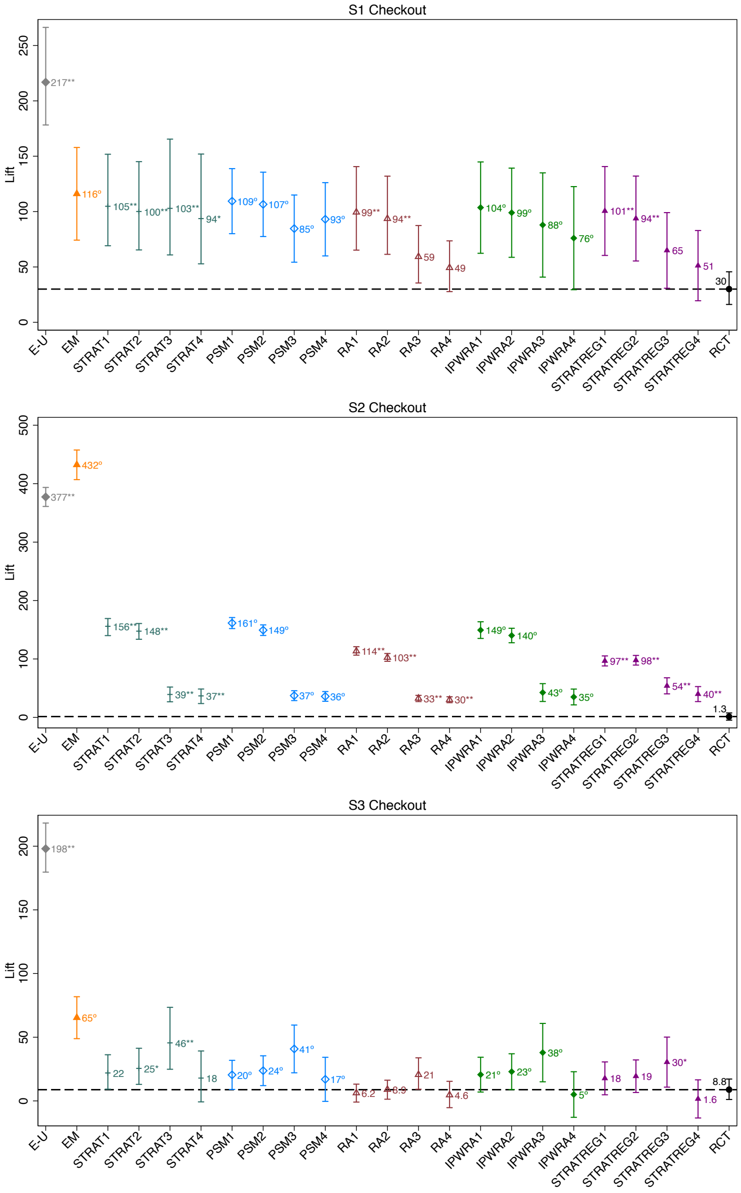


Figure A-8: Checkout conversions: Studies 1, 2, and 3



S1 Checkout

S2 Checkout

S3 Checkout

Figure A-9: Checkout conversions: Studies 4, 5, 7, and 8

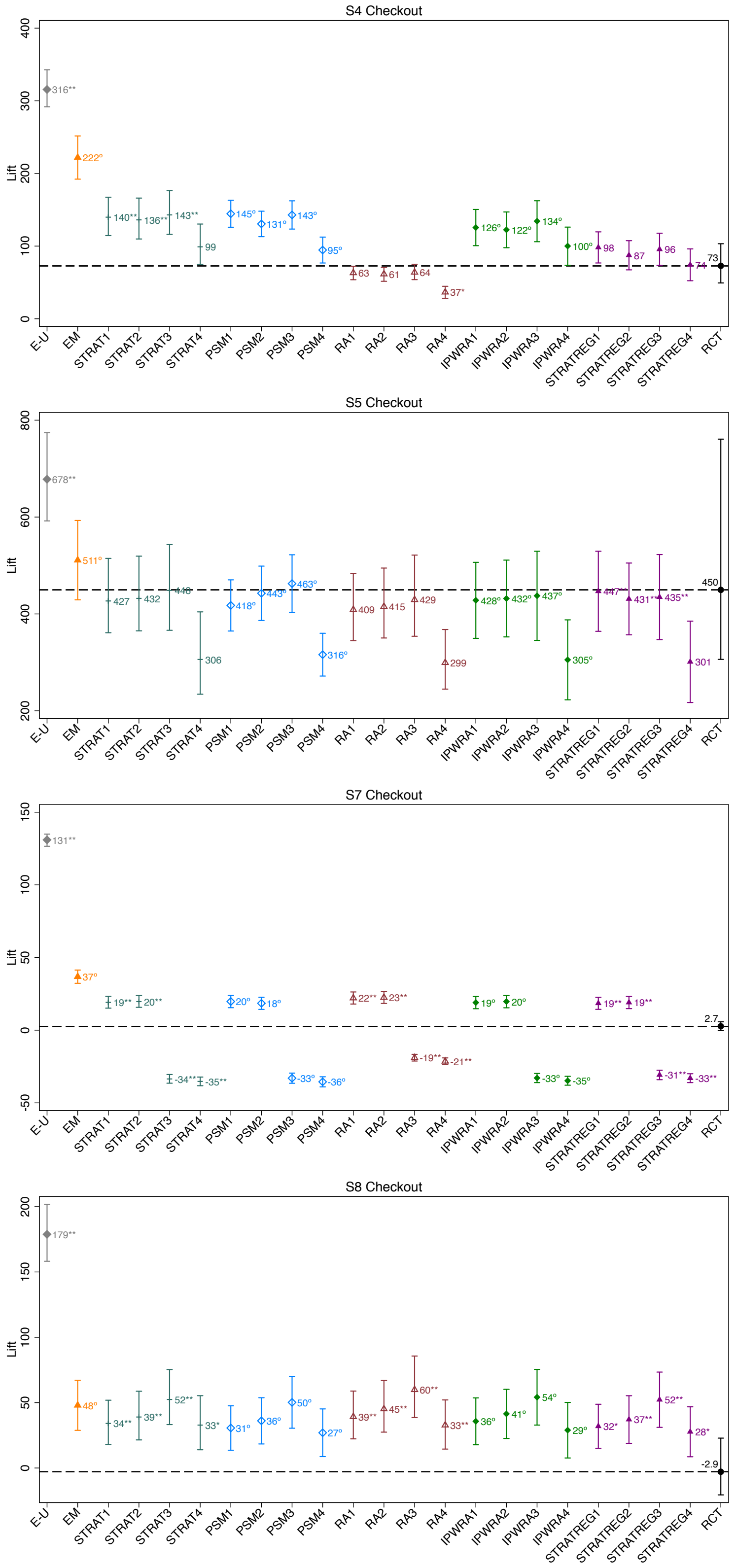


Figure A-10: Checkout conversions: Studies 9, 10, 11, and 12

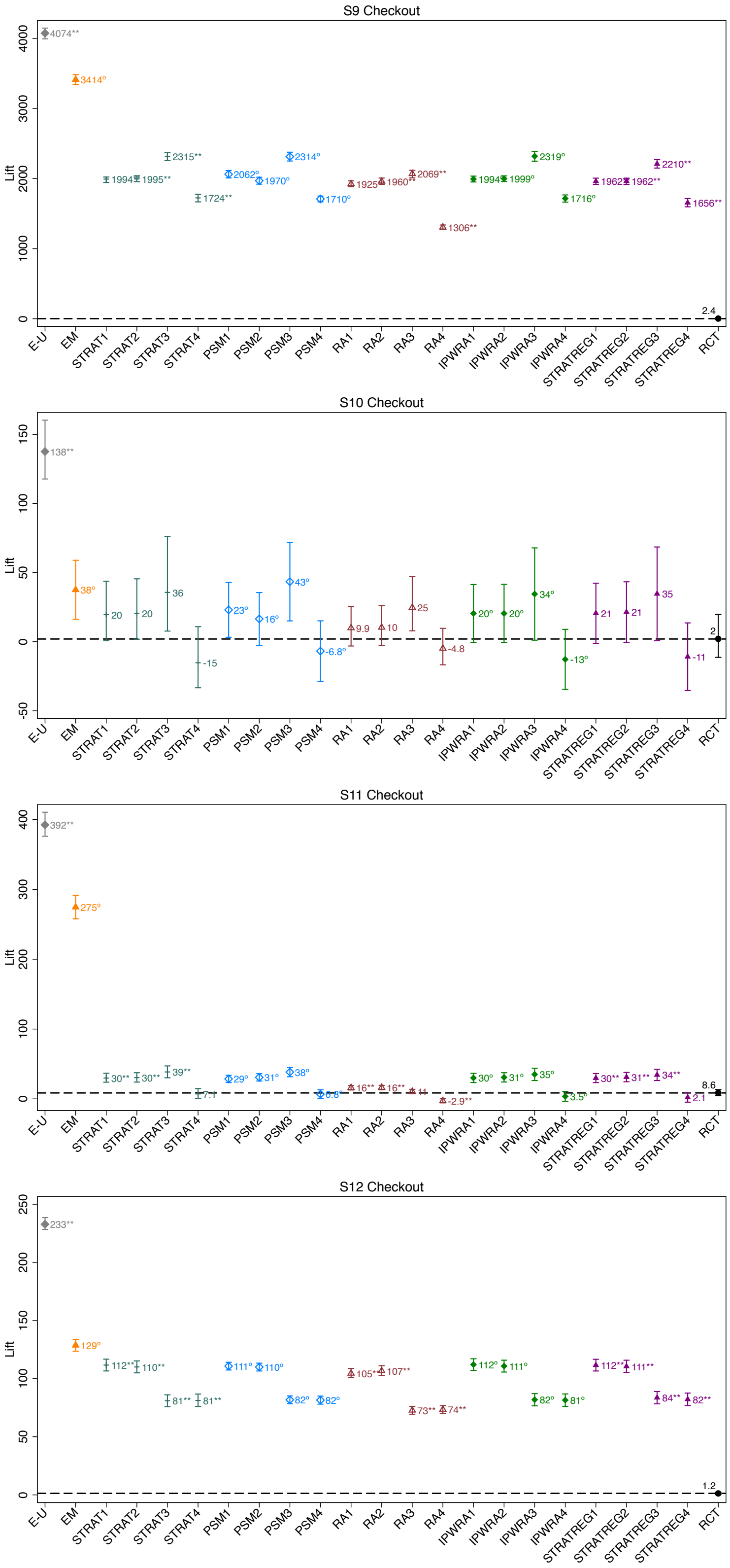


Figure A-11: Checkout conversions: Studies 13, 14, and 15

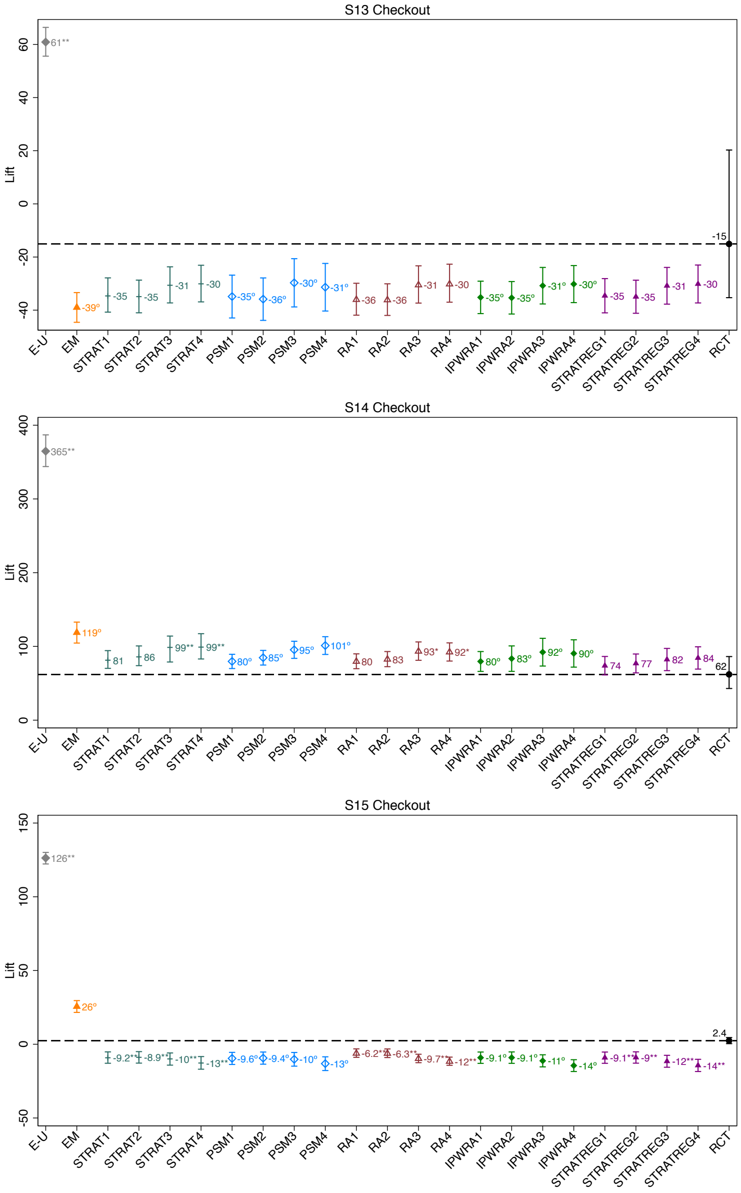


Figure A-12: Registration conversions: Studies 1, 5, 8, and 10

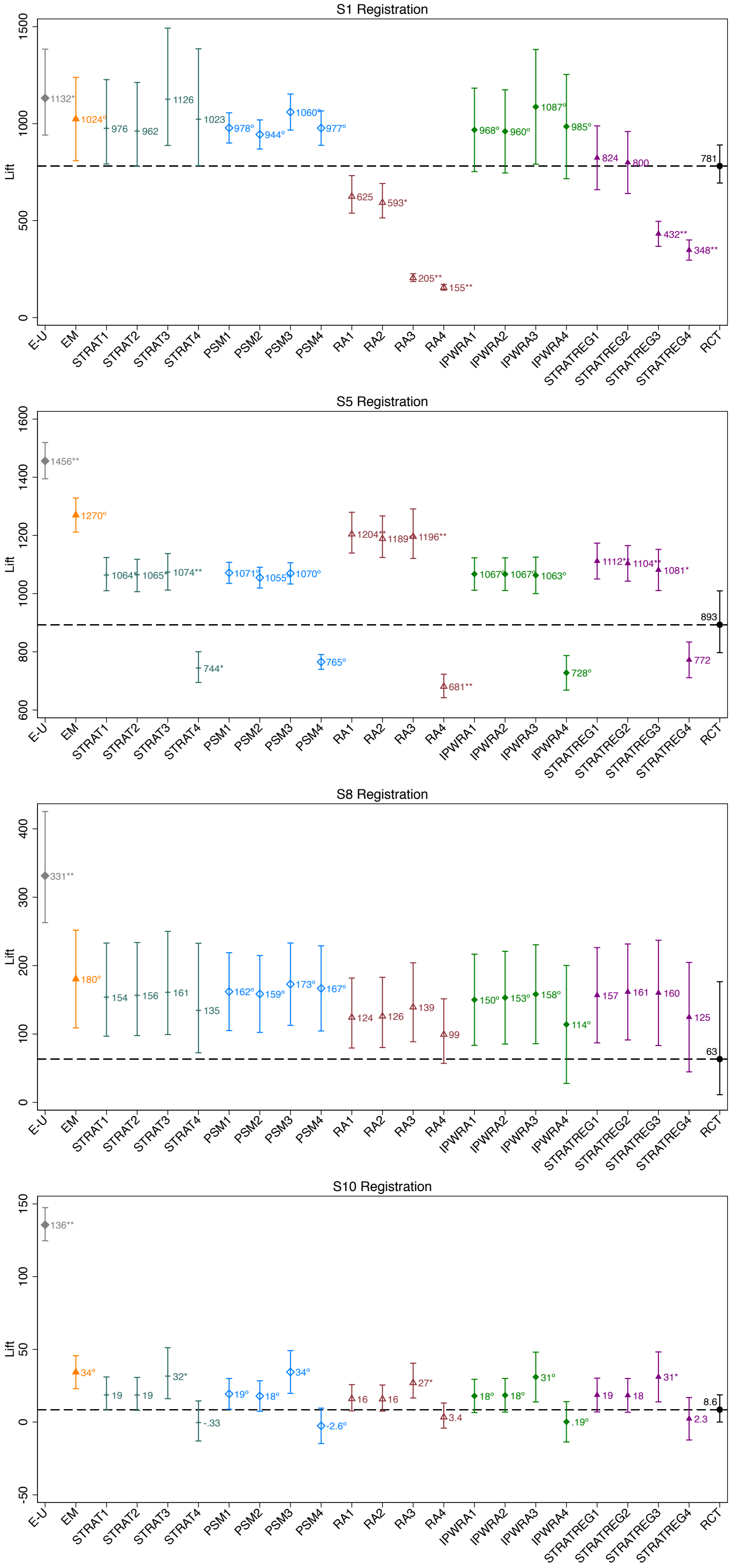
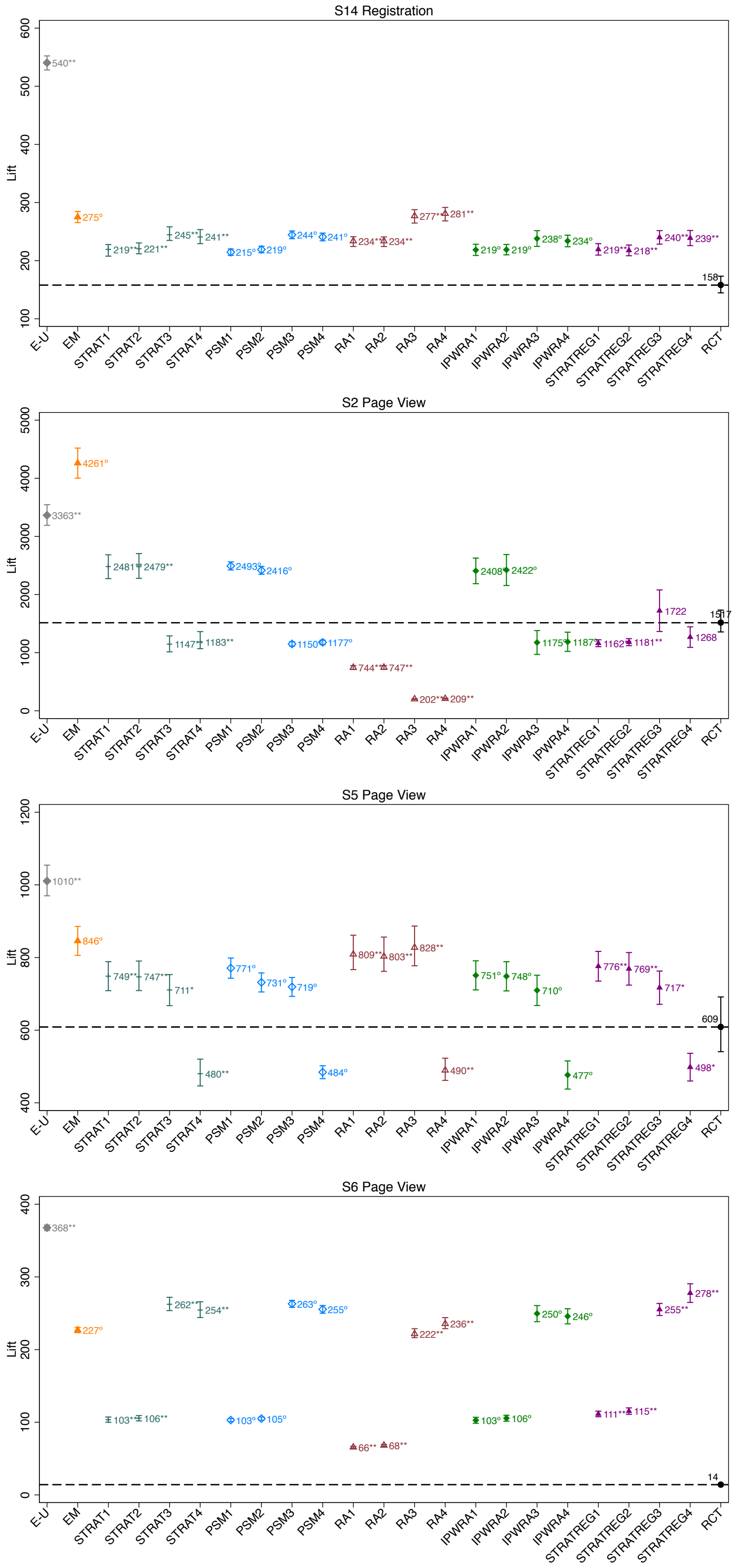


Figure A-13: Registration conversions: Study 14 and page-view conversions: Studies 2, 5, and 6



S14 Registration

S2 Page View

S5 Page View

S6 Page View

Incorporating an Unobservable into an Observational Model

Our approach is based on the methodology developed in Rosenbaum and Rubin (1983a) and extended by Ichino, Fabrizia, and Nannicini (2008). To provide some intuition, we describe the methodology in Rosenbaum and Rubin (1983a). Starting with the modified unconfoundedness assumption in equation (30), assume a binary unobservable, $U \sim Bern(0.5)$, exists correlated with treatment and outcomes.[24] The observational model consists of a logit treatment equation and a linear outcome equation with a normal error term:

$$\Pr(W_i = 1|X_i, U_i) = \frac{\exp(\gamma' X_i + \alpha U_i)}{1 + \exp(\gamma' X_i + \alpha U_i)} \tag{1}$$

$$Y_i = \tau W_i - \beta X_i - \delta U_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \ , \tag{2}$$

where τ is the treatment effect of interest. Note that U_i enters both equations and that the pair of parameters (α, δ) determine the relative importance of U in each equation. Because U is unobserved, we must integrate over it to form the log-likelihood:

$$L(Y, W; \tau, \beta, \gamma, \sigma^2, \alpha, \delta) = \sum_{i=1}^{N} \ln \left[\frac{1}{2} \phi(Y_i - \tau W_i - \beta X_i; \sigma^2) \frac{(\exp(\gamma' X))^{W_i}}{1 + \exp(\gamma' X)} + \right. \tag{3}$$

$$\left. \frac{1}{2} \phi(Y_i - \tau W_i - \beta X_i - \delta; \sigma^2) \frac{(\exp(\gamma' X + \alpha))^{W_i}}{1 + \exp(\gamma' X + \alpha)} \right] \tag{4}$$

where $\phi(\cdot)$ is a Normal density. In practice, optimizing the log-likelihood with respect to (α, δ) is likely difficult because the data are not directly informative of their values. Instead, Rosenbaum and Rubin (1983a) and Imbens (2003) recommend fixing values of (α, δ) and estimating $\theta = (\tau, \beta, \gamma, \sigma^2)$, so that the treatment effect can be expressed as $\tau(\alpha, \delta)$.

In the notation of Rosenbaum and Rubin (1983a), our goal is to find values of (α, δ) such that the treatment effect equals the estimate obtained from the RCT,

$$\widehat{\tau}_{rct} = \tau(\alpha^*, \delta^*) \ . \tag{5}$$

These (α^*, δ^*) characterize the strength of the unobservable U needed to eliminate the bias of the observational method. Note that many values of (α^*, δ^*) might satisfy (5).

The approach above relies on a parametric model for the outcome. To avoid this, we follow Ichino, Fabrizia, and Nannicini (2008), who propose directly specifying the parameters that characterize the distribution of the unobservable U:

$$p_{jk} \equiv \Pr(U_i = 1 | W_i = j, Y_i = k), \qquad j, k \in \{0, 1\}$$

The four parameters $\mathbf{p} = \{p_{00}, p_{01}, p_{10}, p_{11}\}$ define the probability of $U_i = 1$ over each combination of treatment assignment and conversion outcome. We simulate a value of U for each exposed and unexposed user and re-estimate the ATT including this simulated unobservable in the collection of

[24] Although this assumes X and U are independent, the assumption is innocuous because any correlation between them would be accounted for in a suitably flexible observational model. The challenge in estimating treatment effects arises through variation in treatment and outcomes due to the unobservable that cannot be accounted for using the observables.

covariates used to estimate the propensity score. Changing the values of $\mathbf{p}$ produces different types of correlations between the unobservable U and treatment and outcomes.

A close parallel exists between $\mathbf{p}$ and (α, δ). In the Rosenbaum-Rubin sensitivity model, $\delta > 0$ in equation (2) implies that omitting the unobservable U positively biases the estimated treatment effect τ. In Ichino, Fabrizia, and Nannicini (2008), the corresponding direction of bias is achieved by simulating a U with $p_{01} > p_{00}$. To see why, note that measurement bias arises when $Pr(Y_i = 1|W_i = 0, X_i, U_i) \neq Pr(Y_i = 1|W_i = 0, X_i)$, which implies that, without observing U, the outcome of unexposed users cannot be used to estimate the outcome of exposed users in the case of no exposure. This inequality arises when $p_{01} > p_{00}$ because:[25]

$$\begin{aligned} p_{01} > p_{00} &\Rightarrow \Pr(U_i = 1|W_i = 0, Y_i = 1, X_i) > \Pr(U_i = 1|W_i = 0, Y_i = 0, X_i) \\ &\Rightarrow \Pr(Y_i = 1|W_i = 0, U_i = 1, X_i) > \Pr(Y_i = 1|W_i = 0, U_i = 0, X_i) \end{aligned}$$

In the Rosenbaum-Rubin sensitivity model, α determines the strength of the unobservable to lead to selection into treatment. In Ichino, Fabrizia, and Nannicini (2008), the equivalent concept can be found by defining the conditional probability of the treatment assignment $W_i = j$ and the unobservable $U_i = 1$:

$$p_j = \Pr(U_i = 1|W_i = j) = \sum_{k=0}^{1} p_{jk} \cdot \Pr(Y = k|W = j) \ .$$

p_1 is the probability of users drawing an unobservable, $U_i = 1$, conditional on being exposed, $W_j = 1$. Conversely, p_0 is the probability of users drawing an unobservable, $U_i = 1$, conditional on being unexposed, $W_j = 0$. Hence, if we specify that $p_1 > p_0$, a user who draws a positive unobservable is more likely to be exposed rather than unexposed. This corresponds to α in equation (1), or the strength of the unobservable to lead to selection into treatment.

Our sensitivity analysis entails two steps. First, we calculate the ATT while integrating over the unobservable. Recall that adding the outcome model did not improve our estimates significantly. As a result, we estimate the treatment effect by stratification on the propensity score alone. Because we have to compute the treatment effect repeatedly using Monte-Carlo draws, we require a computationally efficient method, which stratification is. First, fix $\{p_{00}, p_{01}, p_{10}, p_{11}\}$ and repeatedly draw R values of $U_i = \{U_i^1, U_i^2, \ldots, U_i^R\}$. Next, for each draw, we calculate an ATT over users and then average over the R ATT estimates. Specifically, for draw $r = 1, \ldots, R$, we follow these steps:

- For each user, draw $U_i^r \sim Bern(p_{jk})$ for $j = W_i$ and $k = Y_i$.
- Estimate the propensity score $\hat{e}_i^r = e(X_i, U_i^r; \hat{\phi}^r)$
- Stratify the estimated propensities $\hat{e}_i^r$ into M bins defined by $B_{im}^r = 1 \cdot \{b_{m-1} < \hat{e}_i^r \leq b_m\}$
- Calculate the ATT across the strata, $\tau^{sen,r}(\mathbf{p})$, as in section 4.2.
- Calculate the average ATT over the draws, $\tau^{sen}(\mathbf{p}) = \frac{1}{R}\sum_r \tau^{sen,r}(\mathbf{p})$.

[25] See the appendix of `http://cepr.org/active/publications/discussion_papers/view_pdf.php?dpno=5736` for a proof.

To explore the full possible range of U that would explain the bias of the estimated ATT, we search over a grid of the probability parameters $\mathbf{p}$. Specifically, we fix $p_{11} = .5$ and pick a pair p_{10} and $p_{01} \in [0, .25, .5, .75, 1]$. We then use Brent's method using $[0, 1]$ as the starting bracket to find a p_{00} such that

$$\widehat{\tau}_{rct} = \tau^{sen}(\mathbf{p}^*) \ .$$

When we cannot find a $p_{00} \in [0, 1]$ given some p_{10} and p_{01} for $p_{11} = .5$, we also search for a solution using $p_{11} = 0$ and $p_{11} = 1$.[26]

Rather than presenting the results in terms of $\mathbf{p}^*$, we take a cue from Imbens (2003) and express the strength of U in terms of the relative variation it explains in treatment assignment and outcomes.[27] This is described in the main body of the text using a slightly less cumbersome, though also less precise, notation than what we use here. We can now redefine these terms using the notation in this appendix, such that the relative strength of the unobservable to affect outcomes is

$$R^2_{Y,rel} = \frac{R^2_Y(\mathbf{p}) - R^2_Y(0)}{R^2_Y(0)}$$

and similar for $R^2_{W,rel}$. That is,

[26] This strategy is informed by the following observation: Suppose that at $p_{11} = .5$ and some p_{10}, p_{01}, a p_{00} exists that solves $\tau^{sen}(\mathbf{p}^*) = \widehat{\tau}_{rct}$. Then, if a p_{00} exists for the same p_{10}, p_{01} but a different p_{11}, p_{00} will be the same as under $p_{11} = .5$ (as will be d and s).

[27] Imbens (2003) characterizes the values for (α, δ) in terms of the share of the unexplained variation in outcomes and treatment, normalizing by $(1 - R^2_Y(0))$ and $(1 - R^2_W(0))$.

Figure A-14: Sensitivity analysis for Studies 1-5, 7-9

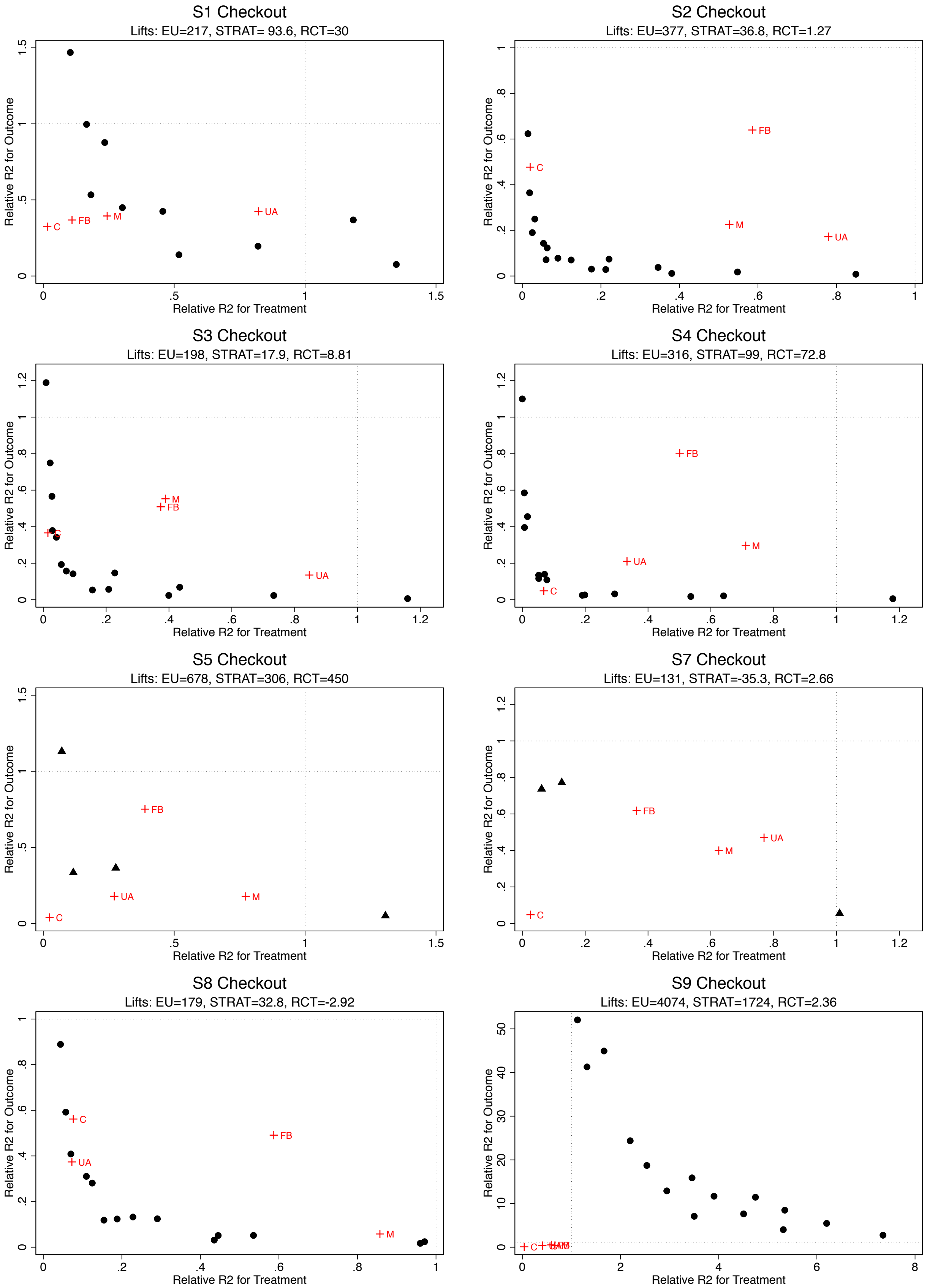
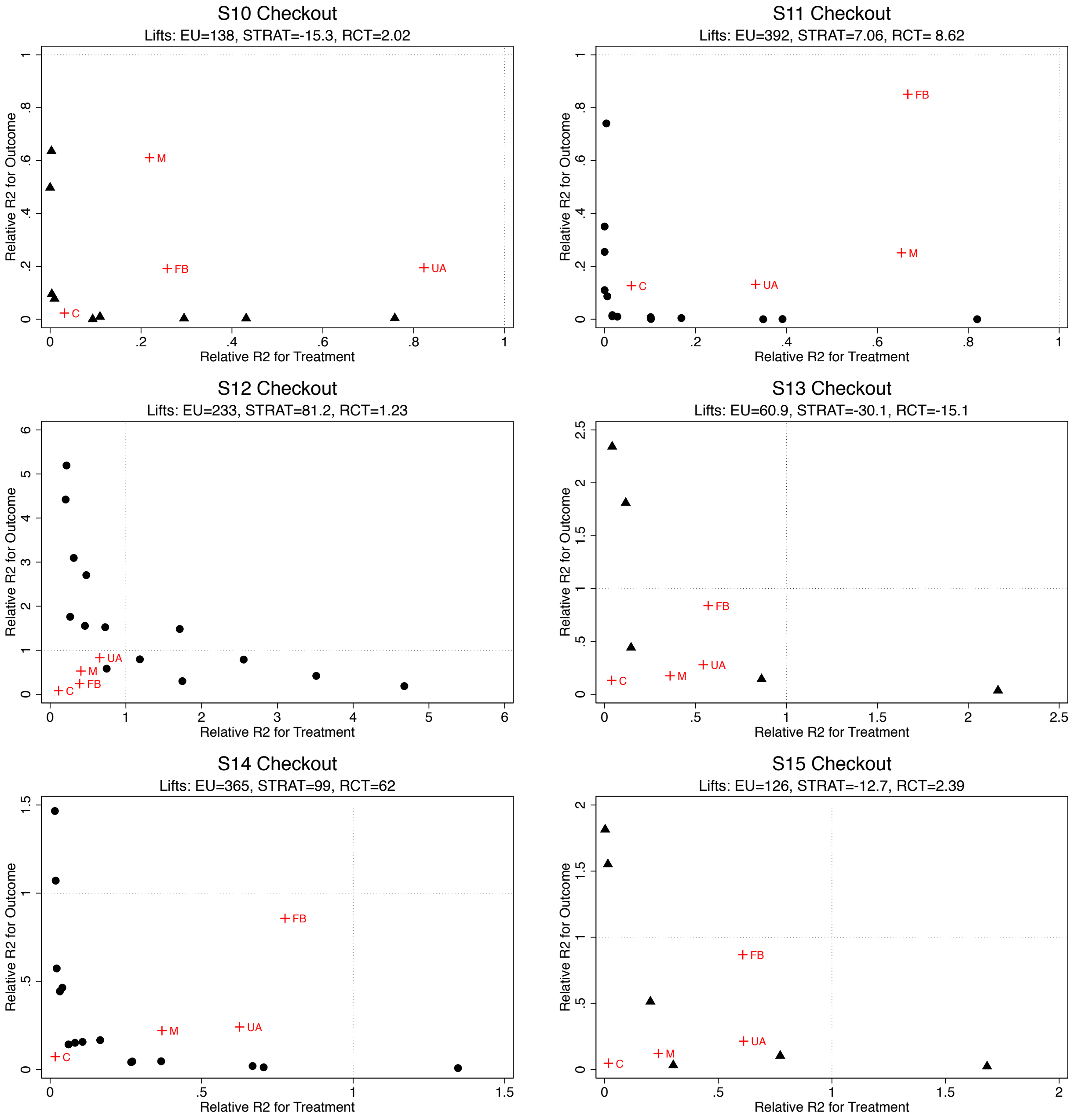


Figure A-15: Sensitivity analysis for Studies 10-15



S10 Checkout

Lifts: EU=138, STRAT=-15.3, RCT=2.02

S11 Checkout

Lifts: EU=392, STRAT=7.06, RCT= 8.62

S12 Checkout

Lifts: EU=233, STRAT=81.2, RCT=1.23

S13 Checkout

Lifts: EU=60.9, STRAT=-30.1, RCT=-15.1

S14 Checkout

Lifts: EU=365, STRAT=99, RCT=62

S15 Checkout

Lifts: EU=126, STRAT=-12.7, RCT=2.39