

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

Brett R. Gordon,^a Florian Zettelmeyer,^{a,b} Neha Bhargava,^c Dan Chapsky^c

^a Kellogg School of Management, Northwestern University, Evanston, Illinois 60208; ^b National Bureau of Economic Research, Cambridge, Massachusetts 02138; ^c Facebook Inc., Menlo Park, California 94025

Contact: b-gordon@kellogg.northwestern.edu,  <http://orcid.org/0000-0001-9081-569X> (BRG); f-zettelmeyer@kellogg.northwestern.edu (FZ); nehab@fb.com (NB); chapsky@fb.com (DC)

Received: November 19, 2017

Revised: April 13, 2018; August 29, 2018

Accepted: September 8, 2018

Published Online in Articles in Advance:

■■■■, 2018

<https://doi.org/10.1287/mksc.2018.1135>

Copyright: © 2018 INFORMS

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. Using data from 15 U.S. 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. In our setting, advances in causal inference methods do not allow us to isolate the exogenous variation needed to estimate the treatment effects. 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.

History: K. Sudhir served as the editor-in-chief and Anja Lambrecht served as associate editor for this article.

Supplemental Material: The online appendix and data files are available at <https://doi.org/10.1287/mksc.2018.1135>.

Keywords: digital advertising • field experiments • causal inference • observational methods • advertising measurement

1. Introduction

Digital advertising spending exceeded television ad spending for the first time in 2017.¹ Advertising is a major 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 et al. 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 data set 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 et al. 2011, Blake et al. 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 predefined window, and variation due to unrelated advertisers' bids. All three forces induce randomness in the ad auction outcomes. However, 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 important attributes that, conditional on the quasi-random variation in ad exposure, 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 that 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 comprehensive individual-level data are “good enough” for ad measurement, or whether even fairly comprehensive

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 fairly comprehensive data prove inadequate to yield reliable estimates of 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 et al. (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) data set 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.

Of course, our results should not be interpreted to suggest that causal inference methods are unable to replicate experimental findings in all settings. For example, Eckles and Bakshy (2017), using Facebook data to measure peer effects (not advertising effects), shows that high-dimensional causal inference models are able to eliminate most of the bias of naïve observational methods. Moreover, these methods may work to recover the causal effects of advertising that is not subject to the same degree of targeting as advertising on Facebook.

We are not the first to attempt to estimate the performance of observational methods in gauging digital advertising effectiveness.⁴ Lewis et al. (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 U.S.-based traffic to the focal ad, leading them to use a matched sample of unexposed international users. Blake et al. (2015) documents that nonexperimental measurement can lead to highly suboptimal spending decisions for online search ads. However, in contrast to our paper, Blake et al. (2015) use an aggregate difference-in-differences approach based on randomization at the level of 210 media markets as the experimental benchmark and therefore did not implement individual-level causal inference methods.

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

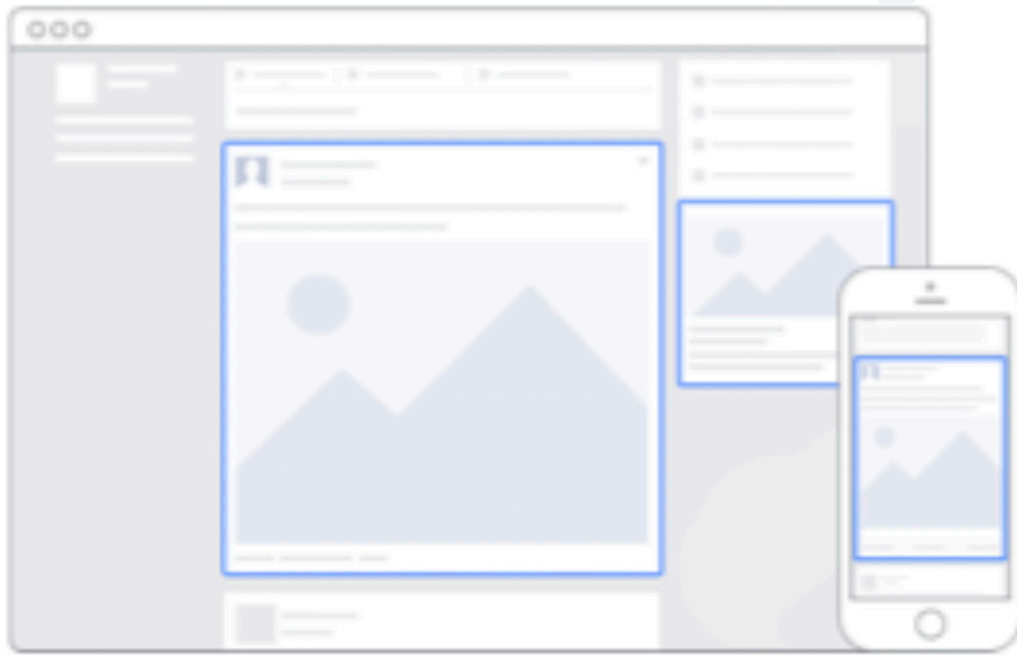
We focus exclusively on campaigns in which the advertiser had a particular “direct response” outcome in mind, for example, to increase sales of a new product.⁶ The industry refers to these as “conversion outcomes.” In each study, the advertiser measured conversion outcomes using a piece of Facebook-provided 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 devices and sessions with a particular user. Such cross-device tracking is important because users are frequently exposed to advertising on a mobile device but might subsequently convert on a tablet or computer.

Figure 1 displays where a Facebook user accessing the site from a desktop/laptop or mobile device might see ads. In the middle is the “News Feed,” where new stories appear with content as the user scrolls down 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 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.

Q: 25, 26 **Figure 1.** (Color online) Facebook Desktop and Mobile-Ad Placement



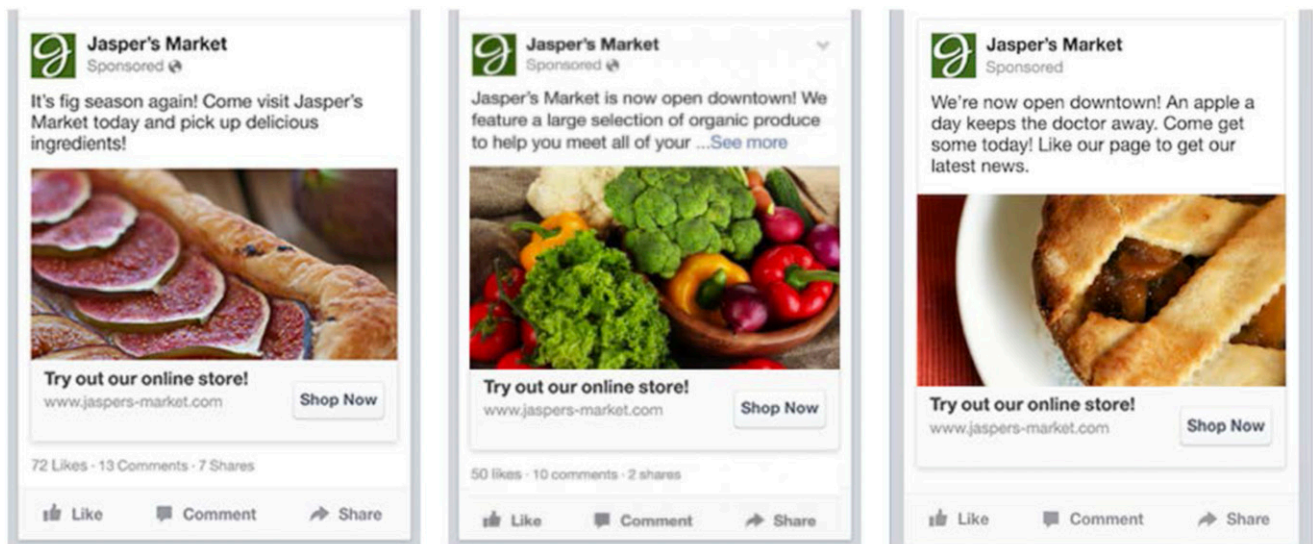
Q: 27 Source. <https://www.facebook.com/business/ads-guide>.

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.

2.2. Experimental Implementation

An experiment begins with the advertiser deciding which consumers to target with a marketing campaign, such as all women between the ages of 18 and 54 years. These targeting rules define the relevant set of users in the study. Each user is randomly assigned to the control or test group on the basis of a proportion selected by the advertiser, in consultation with Facebook. Control-group members are never exposed to campaign

Figure 2. (Color online) Example of Three Display Ads for One Campaign



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

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, owing 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 3(a). 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 3(b) 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.

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 preannounced and occur over relatively short periods. Furthermore, Facebook's scale makes gauging other campaigns' scope or targeting objectives hard for advertisers.

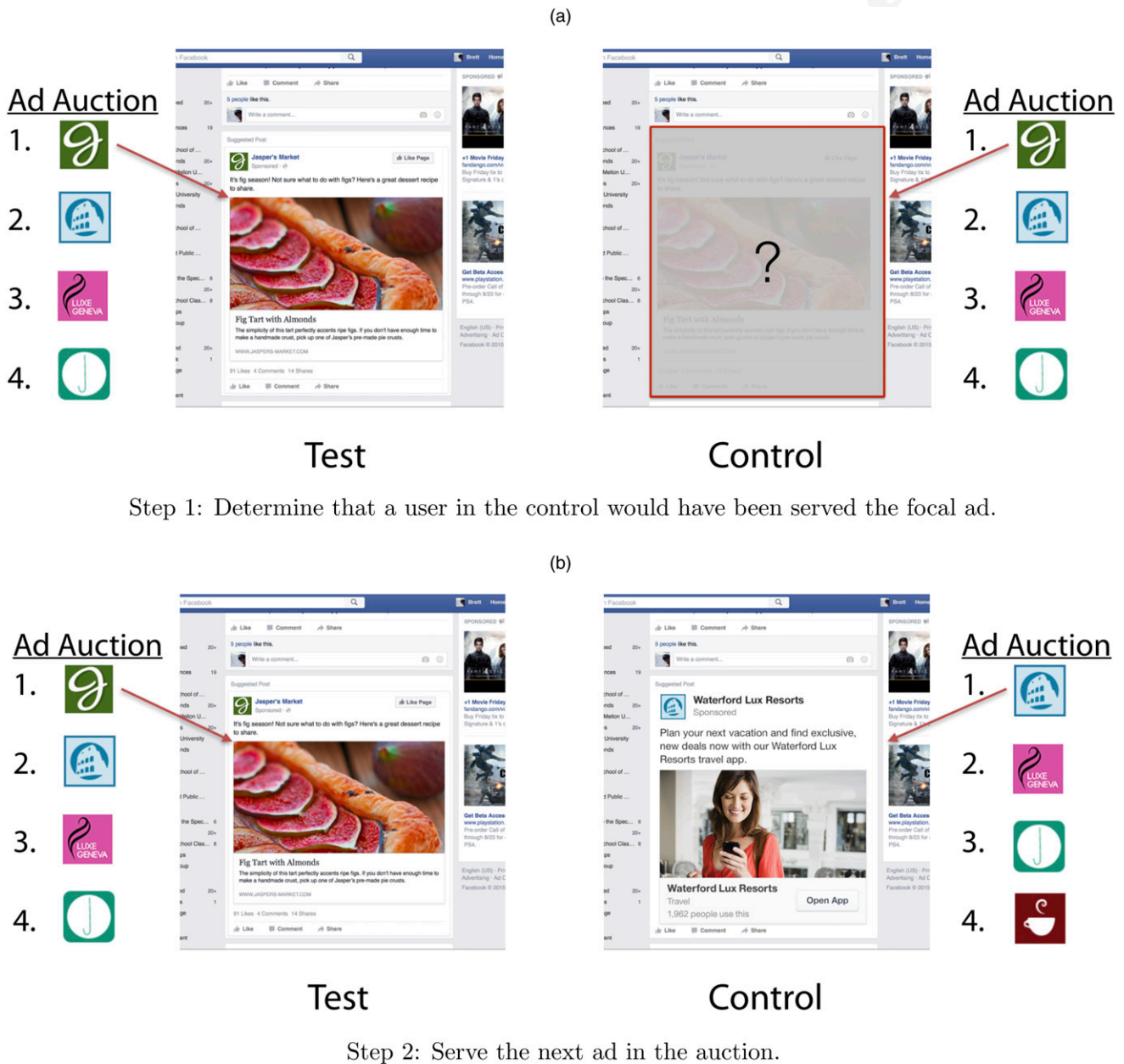
As with any experiment, this one yields an estimate of the campaign's average treatment effect, conditional on all market conditions—such as marketing activities the advertiser conducts in other channels (e.g., search, TV) and its competitors' activities. The estimated lift the experiment yields may 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.

2.3.1. User-Induced Endogeneity. The first mechanism that drives selection was coined "activity bias" when first identified by Lewis et al. (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.

2.3.2. 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

Figure 3. (Color online) Determination of Control Ads in Facebook Experiments

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 on the basis of their higher conversion rates. In general, this mechanism will lead to upwardly biased ad effects, but there are cases in which 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.

2.3.3. 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- to 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 nonexperimental 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 $i = 1, \dots, N$ 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

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

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. Noninterference 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

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, owing to one-sided noncompliance. 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.

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.

When we usually interpret either the ITT or ATT, 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 and ATT should be interpreted as conditional on the platform's ad-optimization system.

3.2. Causal Effects in the RCT

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

$$ITT_Y = \mathbb{E}[Y(1, W(1)) - Y(0, W(0))], \quad (2)$$

with the sample analog being

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

As noted earlier, our focus is on the ATT,

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

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:

$$\tau = \frac{ITT_Y}{ITT_W} = \frac{\mathbb{E}[Y(1, W(1))] - \mathbb{E}[Y(0, W(0))]}{\mathbb{E}[W(1)] - \mathbb{E}[W(0)]}. \quad (5)$$

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 $ITT_W = \mathbb{E}[W(1)]$, 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

$$ITT_Y = ITT_{Y,co} \cdot \pi_{co} + ITT_{Y,nc} \cdot (1 - \pi_{co}), \quad (6)$$

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 $ITT_{Y,nc} = \mathbb{E}[Y(1, 0) - Y(0, 0)] = 0$. Thus, $ITT_{Y,co}$ can be expressed as the ITT effect divided by the share of compliers,

$$\tau \equiv ATT \equiv ITT_{Y,co} = \frac{ITT_Y}{\pi_{co}}. \quad (7)$$

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 noncompliance, 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:

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

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 return on investment (ROI) (Lewis and Rao 2015), but we lack the data on profit margins from sales to calculate ROI.¹⁰

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 use 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.¹¹

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 et al. 2014), measuring the effects of firm-generated social media on customer metrics (Kumar et al. 2016), assessing whether access to digital video recorders affects sales of advertised products (Bronnenberg et al. 2010), the effectiveness of pharmaceutical detailing (Rubin and Waterman 2007), the impact of mutual fund name changes on subsequent investment inflows (Cooper et al. 2005), and evaluating alcohol advertising targeted at adolescents (Ellickson et al. 2005).

In industry, many advertisers rely on third-party measurement vendors, such as comScore, Nielsen, and Nielsen Catalina Solutions, to perform their ad effectiveness measurement. The advertiser may also rely on an advertising agency, which in turn uses the vendor. Vendors use a combination of matching and regression methods to evaluate various marketing programs (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.¹²

Our understanding of the basic approach used in industry is as follows: the vendor has a large database of individual consumers and potentially supplements these data with the advertiser's proprietary data on

consumer demographics and potentially outcomes. The advertising platform provides a set of (hashed) email addresses to the vendor. By applying the same hashing function to its own data, the vendor is able to determine a set of users who were exposed to ads. The vendor then uses the variables contained in its database of individual consumers to identify comparable consumers who were not exposed (the "forensic" controls). Finally, the vendor compares the measured outcome between the exposed and comparable unexposed users.

The variables available to vendors to match vary greatly by vendor and industry. Typically, matches are performed on demographic information about consumers with different levels of sophistication in the matching process. In some cases, depending on the consumer panel of the vendor, variables may include measures of past purchase history and measures of past online activity.

Our approach is to use a sequence of variables that approximate the data that would be available to vendors as inputs into their chosen observational methods (see Section 5 for a detailed description of our data).

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

$$Y_i(W_i) \equiv Y_i(Z_i = 1, W_i). \quad (9)$$

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

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

The ATT obtained using observational method m is

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

and the lift is

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

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

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

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 naïve 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:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i, \quad (14)$$

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,

$$e(x) \equiv \Pr(W_i = 1|X_i = x). \quad (15)$$

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

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

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 used in the literature.

4.2. Observational Methods

4.2.1. Exact Matching. 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 $X_i^{em} \subset \mathbb{X}$. 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

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

The exact matching estimator for the ATT is

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

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

4.2.2. Propensity Score Matching. 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 $m_{i,k}^c$ be the index of the (unexposed) control user that is the k^{th} closest to exposed user i based on $|e(x_{m_{i,k}^c}; \phi) - e(x_i; \phi)|$. The set $\mathcal{M}_i^c = \{m_{i,1}^c, m_{i,2}^c, \dots, m_{i,M}^c\}$ contains the M closest observations for user i . 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

$$\widehat{Y_i(0)} = \frac{1}{M} \sum_{j \in \mathcal{M}_i^c} Y_j^{obs}. \quad (19)$$

The propensity score matching estimator for the ATT is

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

4.2.3. Stratification. 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 $j = 1, \dots, J$. Let B_{ij} be an indicator that user i is contained in stratum j ,

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

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

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

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,

$$\widehat{\tau}^{strat} = \sum_{j=1}^J \frac{N_{1j}}{N_1} \cdot \widehat{\tau}_j^{strat}. \quad (23)$$

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.

4.2.4. Regression Adjustment. 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,

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

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

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

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

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

Note the accuracy of this method depends on how well the covariate distribution for untreated users overlaps the covariate distribution for treated users. Whereas randomized experimentation guarantees such overlap, observational data provide no such guarantee. If the treated users have significantly different observables compared with 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).

4.2.5. Inverse-Probability-Weighted Regression Adjustment. A variant of the RA estimator incorporates information in the propensity scores, borrowing from the insights found in inverse-probability-weighted estimators (Hirano et al. 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-weighted 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, $\hat{\tau}^{ipwra}$. 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).

4.2.6. Stratification and Regression. 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:

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

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

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

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 predefined 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 (BRG and FZ) according to 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.¹³ 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 U.S.-based RCTs, and we restrict attention to users aged 18 years and older.

Table 1 provides summary statistics for each study. The studies range in size, with the smallest containing approximately 2 million users and the largest approximately 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 operating system, 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). These data are only used in our evaluation of the observational

models and are not used by Facebook’s advertising auction system.

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, 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—across all Facebook ad campaigns, not just the campaigns in our sample. This measure 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.

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.

In using this sequence of variable sets, our intention is to approximate the data that would be available to vendors as inputs into their chosen observational methods (with the hope that our data quality exceeds what many vendors have available to them). Clearly,

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	Financial services	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

Note. C, checkout; P, page view; R, registration.

Table 2. Description of Variables

Set	Variable	Description	Source
1	<i>age</i>	Age of user	FB
1	<i>gender</i>	1 = female, 2 = male, 3 = other/unknown	FB
1	<i>rel_status</i>	Married, engaged, in relationship, single, other	FB
1	<i>FB age</i>	Days since user joined FB	FB
1	<i>friends</i>	No. of friends	FB
1	<i>num_initiated</i>	No. of friend requests sent	FB
1	<i>web_L7</i>	No. of last 7 days accessed FB by desktop	FB
1	<i>web_L28</i>	No. of last 28 days accessed FB by desktop	FB
1	<i>mobile_L7</i>	No. of last 7 days accessed FB by mobile	FB
1	<i>mobile_L28</i>	No. of last 28 days accessed FB by mobile	FB
1	<i>mobile_phone_OS</i>	Operating system of primary phone (if exists)	FB
1	<i>tablet_OS</i>	Operating system of primary tablet (if exists)	FB
1	<i>region</i>	Region of user's residence	FB
1	<i>zip</i>	ZIP code of user's residence	FB
2	<i>population</i>	Population in ZIP code	ACS
2	<i>housingunits</i>	No. of housing units	ACS
2	<i>pctblack</i>	% Black residences	ACS
2	<i>pctasian</i>	% Asian residences	ACS
2	<i>pctwhite</i>	% White residences	ACS
2	<i>pcthispanic</i>	% Hispanic residences	ACS
2	<i>pctunder18</i>	% Residents under age 18 years	ACS
2	<i>pctmarriedhh</i>	% Married households	ACS
2	<i>yearbuilt</i>	Average year residences built	ACS
2	<i>pcths</i>	% Residents with at most high school degree	ACS
2	<i>pctcol</i>	% Residents with at most college degree	ACS
2	<i>pctgrad</i>	% Residents with graduate degree	ACS
2	<i>pctbusfinance</i>	% Working in business/finance	ACS
2	<i>pctstem</i>	% Working in STEM	ACS
2	<i>pctprofessional</i>	% Working in professional jobs	ACS
2	<i>pcthealth</i>	% Working in health industry	ACS
2	<i>pctprotective</i>	% Working in protective services	ACS
2	<i>pctfood</i>	% Working in food industry	ACS
2	<i>pctmaintenance</i>	% Working in maintenance	ACS
2	<i>pcthousework</i>	% Working in home services	ACS
2	<i>pctsales</i>	% Working in sales	ACS
2	<i>pctadmin</i>	% Working in administration	ACS
2	<i>pctfarmfish</i>	% Working at farms or fisheries	ACS
2	<i>pctconstruction</i>	% Working in construction	ACS
2	<i>pctrepair</i>	% Working in repair industry	ACS
2	<i>pctproduction</i>	% Working in production industry	ACS
2	<i>pcttransportation</i>	% Working in transportation industry	ACS
2	<i>income</i>	Average household income	ACS
2	<i>medhhsz</i>	Median household size	ACS
2	<i>medhval</i>	Median household value	ACS
2	<i>vehperh</i>	Average vehicles per household	ACS
2	<i>pctowned</i>	% Households who own a home	ACS
2	<i>pctvacant</i>	% Vacant residences	ACS
2	<i>pctunemployed</i>	% Unemployment	ACS
2	<i>pctlimenglish</i>	% Residents with limited English	ACS
2	<i>pctpoverty</i>	% Residents living below poverty line	ACS
3	<i>mobile_activity</i>	Decile of users' FB activity on mobile devices	FB
3	<i>desktop_activity</i>	Decile of users' FB activity on desktop devices	FB
3	<i>other_activity</i>	Decile of users' FB activity on other devices	FB
4	<i>match_score</i>	Composite variable of FB data	FB

Notes. First three rows are self-reported by the users. Region and ZIP code are determined by geo-location. ACS data are from 2010. FB, Facebook; STEM, science, technology, engineering, and medicine.

our approach has its limitations in that we do not observe precampaign outcome histories that vendors might have received from an advertiser. Therefore, one can interpret our exercise as being most comparable to

what an advertiser would have at their disposal for new customer acquisition.

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 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, 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.

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. After using the observables to control for the endogenous variation in selection, one risk is that there may be little remaining variation with which to estimate the treatment effect.

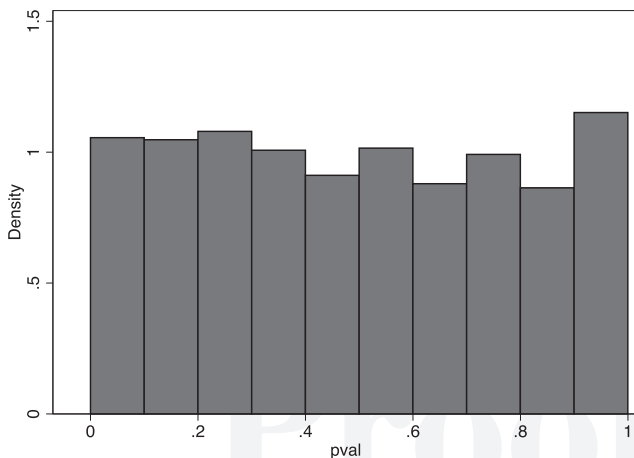
6.2. Estimation and Inference

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

Figure 4. Distribution of p -Values Across All Studies



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 et al. (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, stratification with regression [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

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:

$$t = \frac{\tau_\ell - \tau_\ell^m}{\sqrt{\text{Var}(\tau_\ell) + \text{Var}(\tau_\ell^m) - 2\text{Cov}(\tau_\ell, \tau_\ell^m)}}. \quad (29)$$

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 $\text{Cov}(\tau_\ell, \tau_\ell^m)$ 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 et al. (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 $\tilde{X} = \tilde{X}^W \cup \tilde{X}^Y$.

Step 2: Analysis Using the Bootstrap. For $s = 1, 2, \dots, S$, 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 .

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_ℓ^m .

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

6.3. Assessing Balance

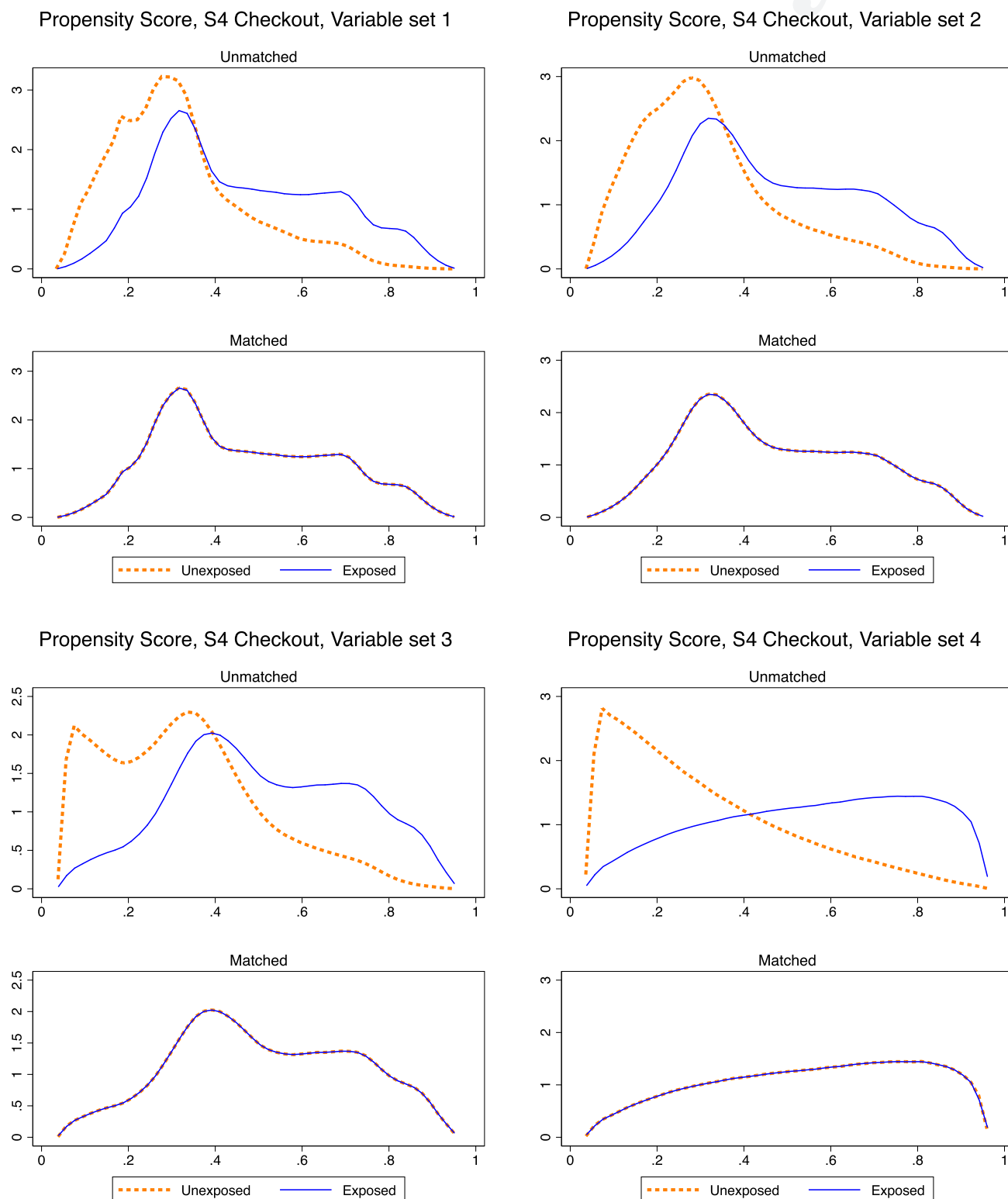
The key assumption for all the observational methods is unconfoundedness, which implies treatment is independent of potential outcomes after conditioning on observables. Rosenbaum and Rubin (1983b) show that unconfoundedness conditional on the observables implies unconfoundedness conditional on the propensity 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

Figure 5. (Color online) Study 4 Density of Estimated Propensity Scores by Treatment and Pre/Post Matching



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 owing 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 seems to remain in the difference of the propensity scores between the exposed and unexposed group.

Detailed graphs with the results of all remaining studies can be found in the online appendix, starting on page OA-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.

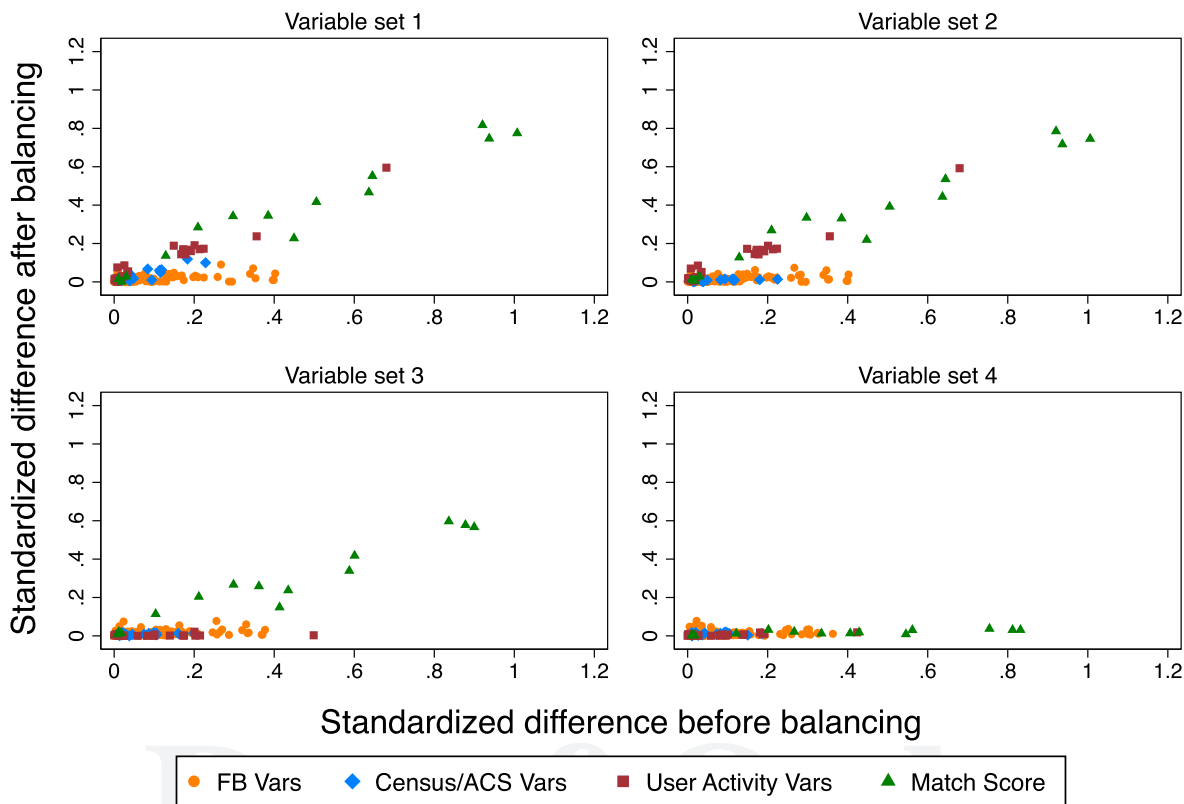
Given that we achieve balance on the propensity score distributions, how well does this balance the actual observables? Following the recommendations of Gelman and Hill (2007) and Imbens and Wooldridge (2009), we examine the standardized differences in covariate means between exposed and unexposed groups. Before matching, we expect to observe differences in the covariate means between the exposed and unexposed group because selection into exposures is nonrandom.

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 Section 5 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 approximately 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 1 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

Figure 6. (Color online) Absolute Standardized Differences of Covariate Means for Study 4



Note. The sets of variables correspond to those described on page 21.

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 can be found in the online appendix, starting on page OA-4. Balancing on the propensity score substantially reduces the mean differences in all the studies and achieves good balance on the underlying observables.

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. Randomized Controlled Trial

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

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 percentage of consumers who were exposed to the ad (37%), yielding an ATT of 0.033%. On the basis of 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

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 with 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

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

Figure 7. (Color online) Results from RCT

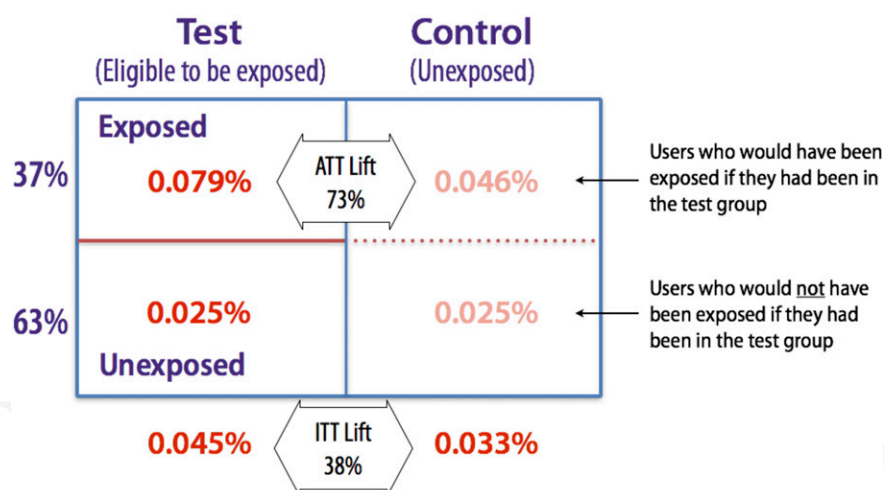


Table 3. ITT Lift for All Studies and Measured Outcomes

Study	Outcome	Conversion probability (%)		RCT ITT (%)	RCT ITT lift (%)	RCT ITT lift confidence interval (%)
		Control group	Test group			
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]

Notes. RCT ITT and RCT ITT lift in bold indicate statistically different from zero at the 5% level. The 95% confidence intervals for RCT ITT lift obtained via bootstrap.

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.

Table 4. ATT Lift for All Studies and Measured Outcomes

Study	Outcome	Exposed (%)	Conversion probability (%)		RCT ATT (%)	RCT ATT lift (%)	RCT ATT lift confidence interval (%)
			Exposed in test	Unexposed in test			
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	1,517.1	[1,357, 1,733]
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]

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

Table 5. Mean by Exposure

Variable	Unexposed	Exposed
Age	26.4	29.3
Female	88%	96%
Facebook age	2,202	2,242
No. 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%

For example, the second item there shows that exposed users are approximately eight percentage points more likely to be female than unexposed users. The table also demonstrates that, compared with 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.

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%.

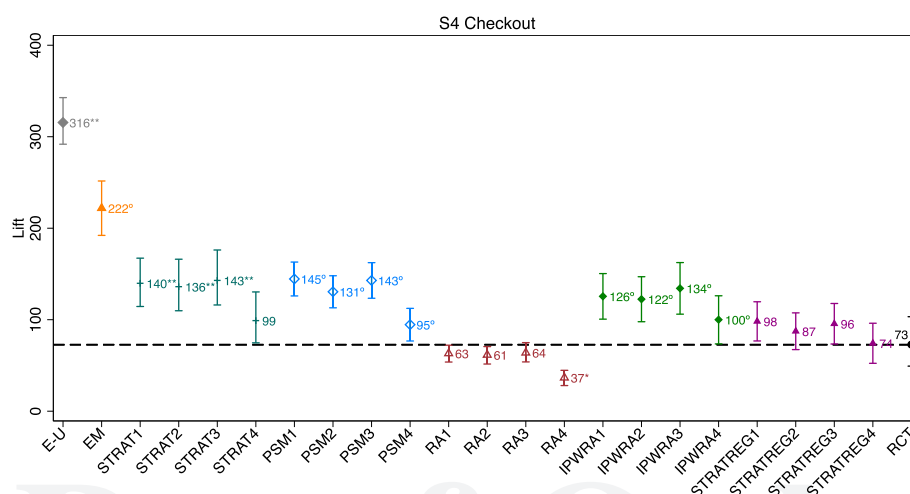
In the remainder of this section, we use multiple observational methods to estimate advertising effectiveness. Our goal is to assess how close these estimates come to the RCT ATT lift benchmark.

We first present three methods that match exposed and unexposed groups but do not rely on an outcome model: (1) exact matching based only on age and gender (as a naive benchmark used widely in industry), (2) stratification (STRAT), and (3) propensity score matching. 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) 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.

7.2.1. 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 propensity score matching).

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 on age and gender alone performs poorly, yielding a lift of 222%.

Figure 8. (Color online) Summary of Lift Estimates and Confidence Intervals for Study 4



Notes. *We reject the hypothesis at a 5% significance level. **We reject the hypothesis at a 1% significance level. []We fail to reject the hypothesis. °We cannot draw an inference.

All methods (except for the regression model, RA) show the following pattern: variable sets 1–3 overestimate lift compared with the RCT by approximately the same amount. The richest set of explanatory variables, variable set 4, yields 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?

7.2.2. Study 1. Study 1's results follow a similar pattern to those in study 4 (see top panel in Figure 9). 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.

7.2.3. 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 1,306% (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.

7.2.4. 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.

7.3. Summary of All 15 Studies

Detailed graphs with the results of all remaining studies can be found in the online appendix, starting on page OA-9. 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 three or more. The color is proportional to the magnitude of misestimation.

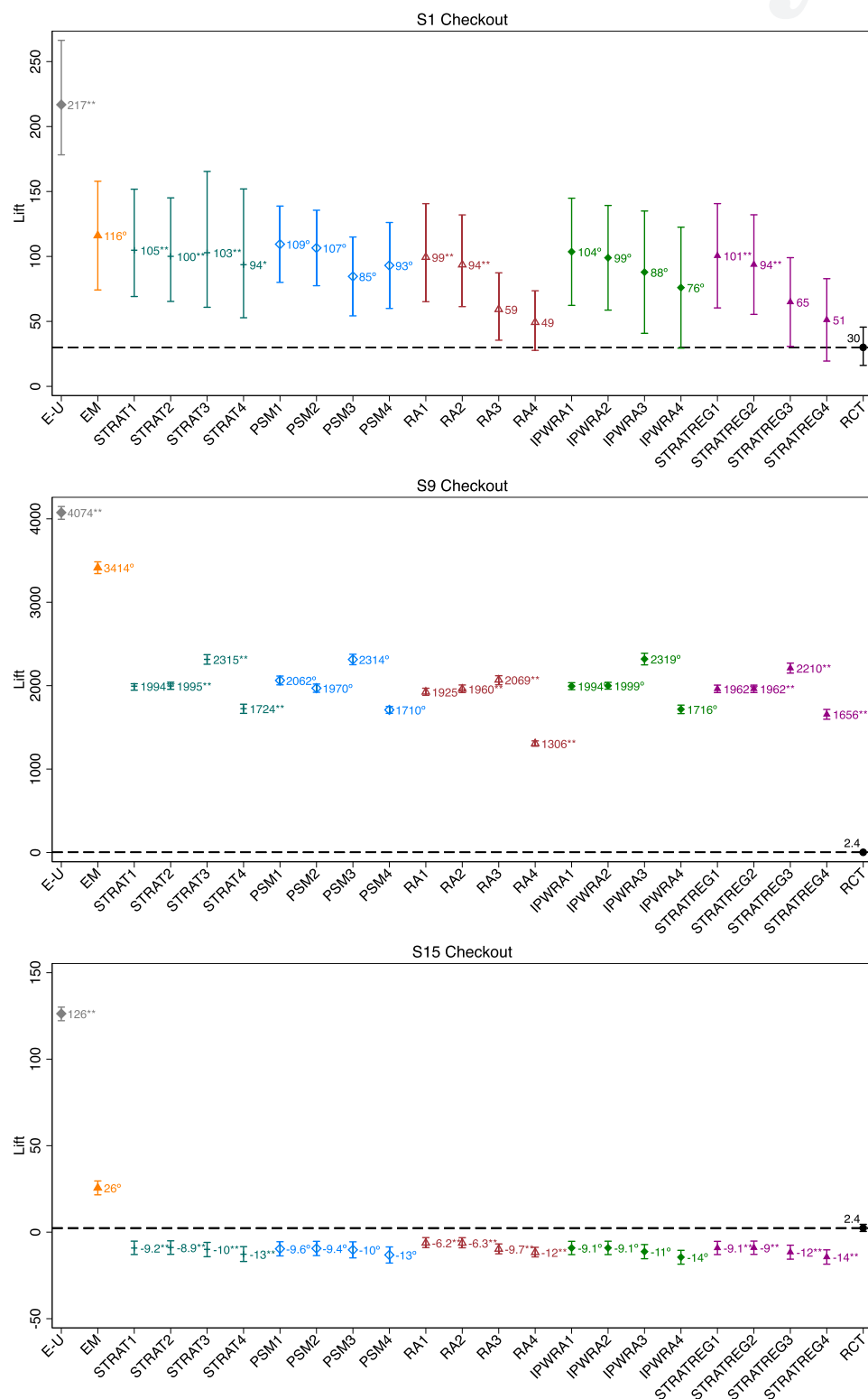
A scan of Figure 10 reveals several clear patterns. First, the observational methods we study mostly overestimate the RCT lift, although in some cases they can significantly underestimate RCT lift. Second, the point estimates in 7 of the 14 studies with a checkout–conversion outcome are consistently 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.

Figure 9. (Color online) Summary of Lift Estimates and Confidence Intervals



Notes. *We reject the hypothesis at a 5% significance level. **We reject the hypothesis at a 1% significance level. []We fail to reject the hypothesis. °We cannot draw an inference.

Q: 30, 32 Figure 10. (Color online) Summary of Lift Results

	Study	RCT Lift	E-U		Stratification				Propensity Score Matching				Regression				Inv Prob. Weighted Regression Adjustment				Stratified Regression			
			Age, Gender	Age, Gender + FB	Age, Gender + FB + Censur	Age, Gender + FB + Censur + Activity	Age, Gender + FB + Censur + Activity + FB Match	Age, Gender + FB	Age, Gender + FB + Censur	Age, Gender + FB + Censur + Activity	Age, Gender + FB + Censur + Activity + FB Match	Age, Gender + FB	Age, Gender + FB + Censur	Age, Gender + FB + Censur + Activity	Age, Gender + FB + Censur + Activity + FB Match	Age, Gender + FB	Age, Gender + FB + Censur	Age, Gender + FB + Censur + Activity	Age, Gender + FB + Censur + Activity + FB Match	Age, Gender + FB	Age, Gender + FB + Censur	Age, Gender + FB + Censur + Activity	Age, Gender + FB + Censur + Activity + FB Match	
Checkout																								
1		30%	217%	116%	105%	100%	103%	94%	109%	107%	85%	93%	99%	94%	59%	49%	104%	99%	88%	76%	101%	94%	65%	51%
2		1.3%	377%	432%	156%	148%	39%	37%	161%	149%	37%	36%	114%	103%	33%	30%	149%	140%	43%	35%	97%	98%	54%	40%
3		8.8%	198%	65%	22%	25%	46%	18%	20%	24%	41%	17%	6%	9%	21%	5%	21%	23%	38%	5%	18%	19%	30%	2%
4		73%	316%	222%	140%	136%	143%	99%	145%	131%	143%	95%	63%	61%	64%	37%	126%	122%	134%	100%	98%	87%	96%	74%
5		450%	678%	511%	427%	432%	448%	306%	418%	443%	463%	316%	409%	415%	429%	299%	428%	432%	437%	305%	447%	431%	435%	301%
7		2.7%	131%	37%	19%	20%	-34%	-35%	20%	18%	-33%	-36%	22%	23%	-19%	-21%	19%	20%	-33%	-35%	19%	19%	-31%	-33%
8		-2.9%	179%	48%	34%	39%	52%	33%	31%	36%	50%	27%	39%	45%	60%	33%	36%	41%	54%	29%	32%	37%	52%	28%
9		2.4%	4074%	3414%	1994%	1995%	2315%	1724%	2062%	1970%	2314%	1710%	1925%	1960%	2069%	1306%	1994%	1999%	2319%	1716%	1962%	1962%	2210%	1656%
10		2.0%	138%	38%	20%	20%	36%	-15%	23%	16%	43%	-7%	10%	10%	25%	-5%	20%	20%	34%	-13%	21%	21%	35%	-11%
11		9%	392%	275%	30%	30%	39%	7%	29%	31%	38%	7%	16%	16%	11%	-3%	30%	31%	35%	3%	30%	31%	34%	2%
12		1%	233%	129%	112%	110%	81%	81%	111%	110%	82%	82%	105%	107%	73%	74%	112%	111%	82%	81%	112%	111%	84%	82%
13		-15%	61%	-39%	-35%	-35%	-31%	-30%	-35%	-36%	-30%	-31%	-36%	-31%	-30%	-35%	-35%	-31%	-30%	-35%	-35%	-31%	-30%	
14		62%	365%	119%	81%	86%	99%	99%	80%	85%	95%	101%	80%	83%	93%	92%	80%	83%	92%	90%	74%	77%	82%	84%
15		2%	126%	26%	-9%	-9%	-10%	-13%	-10%	-9%	-10%	-13%	-6%	-6%	-10%	-12%	-9%	-9%	-11%	-14%	-9%	-9%	-12%	-14%
Registration																								
1		781%	1132%	1024%	976%	962%	1126%	1023%	978%	944%	1060%	977%	625%	593%	205%	155%	968%	960%	1087%	985%	824%	800%	432%	348%
5		893%	1456%	1270%	1064%	1065%	1074%	744%	1071%	1055%	1070%	765%	1204%	1189%	1196%	681%	1067%	1067%	1063%	728%	1112%	1104%	1081%	772%
8		63%	331%	180%	154%	156%	161%	135%	162%	159%	173%	167%	124%	126%	139%	99%	150%	153%	158%	114%	157%	161%	160%	125%
10		9%	136%	34%	19%	19%	32%	0%	19%	18%	34%	-3%	16%	16%	27%	3%	18%	18%	31%	0%	19%	18%	31%	2%
14		158.1%	540%	275%	219%	221%	245%	241%	215%	219%	244%	241%	234%	234%	277%	281%	219%	219%	238%	234%	219%	218%	240%	239%
Page View																								
2		1517%	3363%	4261%	2481%	2479%	1147%	1183%	2493%	2416%	1150%	1177%	744%	747%	202%	209%	2408%	2422%	1175%	1187%	1162%	1181%	1722%	1268%
5		609%	1010%	846%	749%	747%	711%	480%	771%	731%	719%	484%	809%	803%	828%	490%	751%	748%	710%	477%	776%	769%	717%	498%
6		14%	368%	227%	103%	106%	262%	254%	103%	105%	263%	255%	66%	68%	222%	236%	103%	106%	250%	246%	111%	115%	255%	278%

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

Red: Observational method overestimates lift

Blue: Observational method underestimates lift

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

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

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

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,

$$(Y_i(0), Y_i(1)) \perp W_i | X_i, U_i. \quad (30)$$

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 et al. (2008), is to simulate an unobservable U that eliminates this bias.²³ We direct the reader to the online appendix for details.

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 $R_Y^2(X, U)$ be the R^2 from a regression of the outcome on (X, U) for the subset of observations with $W_i = 0$. Similarly, $R_Y^2(X) \equiv R_Y^2(X, 0)$ is the R^2 from a regression on X alone for untreated users. The relative strength of the unobservable to affect outcomes is

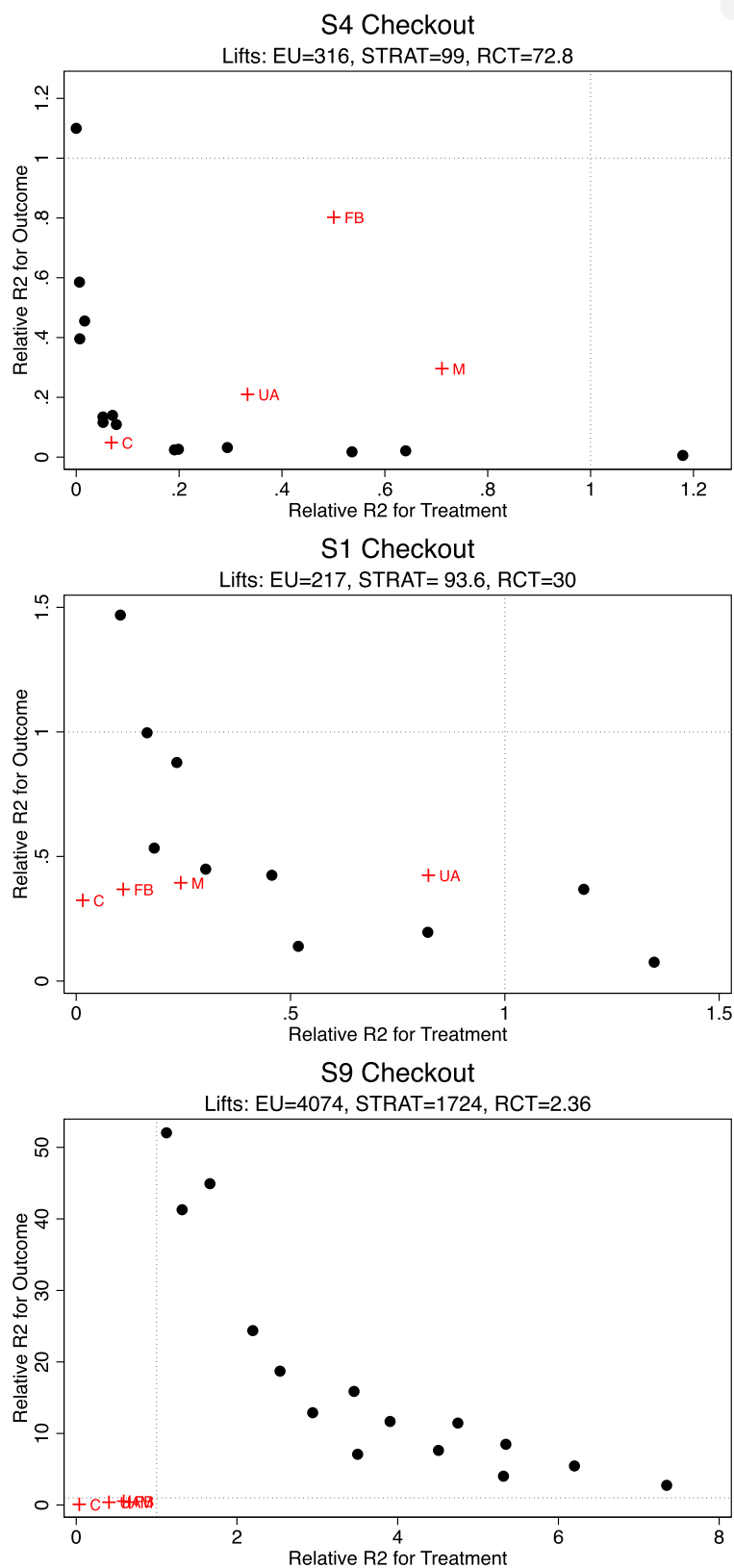
$$R_{Y,rel}^2 = \frac{R_Y^2(X, U) - R_Y^2(X)}{R_Y^2(X)}$$

and similar for $R_{W,rel}^2$. The interpretation is as follows: $R_{Y,rel}^2 = 1$, and $R_{W,rel}^2 = 1$ 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, $R_{Y,rel}^2 = 2$, and $R_{W,rel}^2 = 3$, 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%, compared with 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 variables alone explain approximately 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 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

Figure 11. (Color online) Sensitivity Analysis for Studies 4, 1, and 9



Note. FB, FB variables; C, Census variables; UA, user-activity variables; M, match-score variables.

remove the bias entirely. The unobservable could be the same strength as the census data, which explains approximately 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 approximately 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, 1,724% relative to the 2.36%, with a combined treatment and selection effect of 4,047%. 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 can be found in the Appendix. 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:

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

Table 6 sorts the studies by this ratio. Visual inspection of the graphs in Figure A.1 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, 15, and 10. Note that little additional information is a statement that is relative to the bias reduction we have already achieved using observational data. We caution that the remaining bias may be very hard to eliminate. For example, although we have already eliminated 90% of the bias in study 2 ($brr = 10\%$), we need data that would allow us to reduce the remaining bias from 37% to 1.3%. It may very well be that the “last mile” is the hardest. Studies in which one would need massive additional information relative to what we observe are 12, 9, and 5. Studies 8, 13, 7, and 1 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 variable mostly explains treatment exposure rather than outcome. This is consistent with the finding in Lewis et al. (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 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.

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	4,074	1,724	2.4	42
5	Checkout	678	306	450	63

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 over-estimates 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.

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, advertisers often have private information, from their own businesses or through third-party data providers, to inform their bidding strategies.

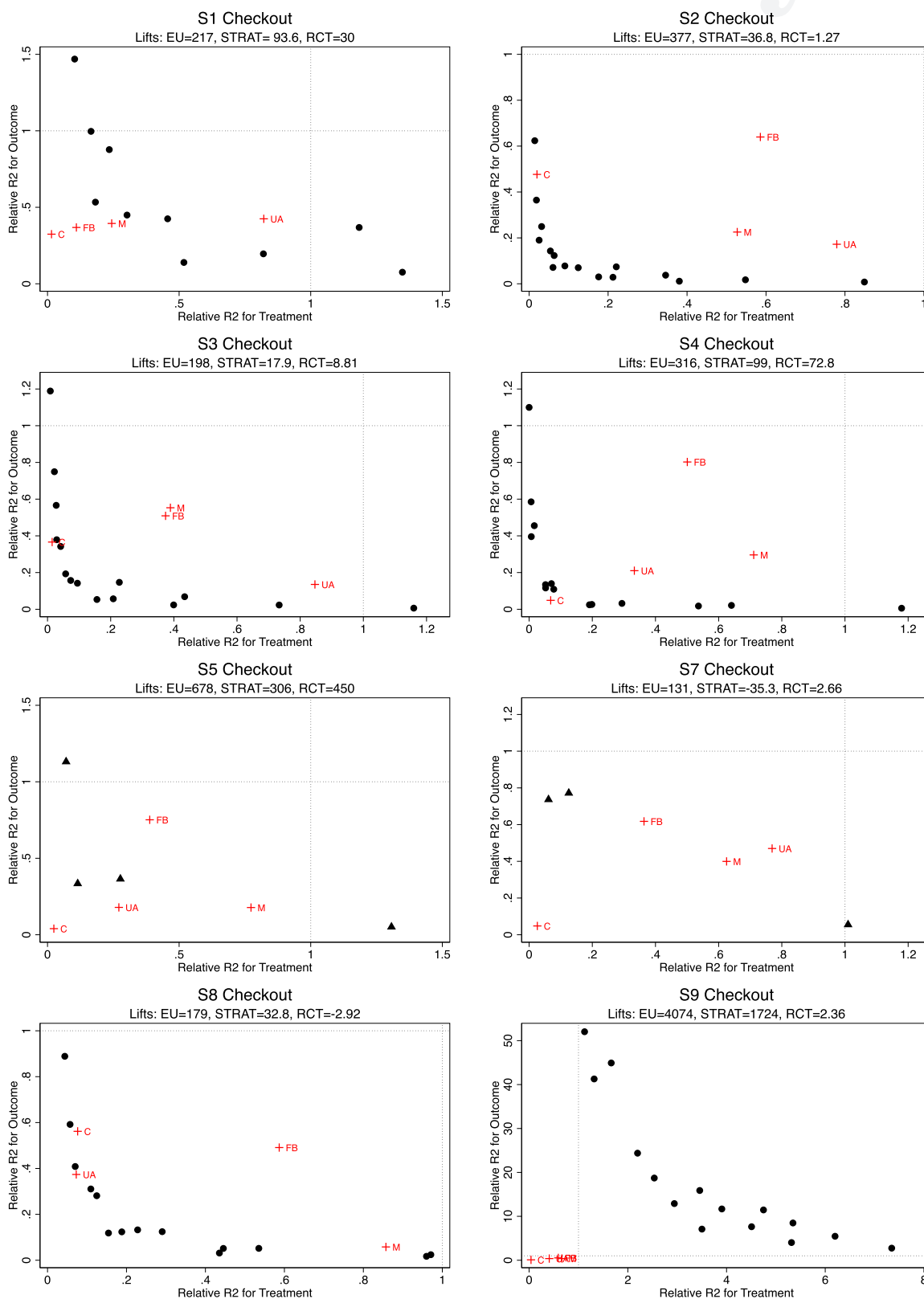
One caveat related to our conclusion is that the performance of the observational methods we study is only as good as the data and the variation they contain. In particular, the failure of observational methods is specific to the Facebook data we had at our disposal. It did not encompass all the data Facebook relies on in its ad-delivery system, nor does Facebook necessarily log or retain all these data for future analysis. Moreover, performing this analysis on a different advertising platform may yield different results.

Acknowledgments

The authors thank Daniel Slotwiner, Robert Moakler, Gabrielle Gibbs, Joseph Davin, Brian d'Alessandro, and Fangfang Tan at Facebook; 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, Stanford, and ZEW for helpful comments and suggestions; and in particular Meghan Busse for extensive comments and editing suggestions. BRG and FZ were not compensated by Facebook or its affiliated companies for engaging in this research. To facilitate data access for future projects, BRG and FZ are currently contingent employees of Facebook. They receive a token salary of \$15 per week, which they donate to charity. To maintain privacy, no data contained personally identifiable information that could identify consumers or advertisers.

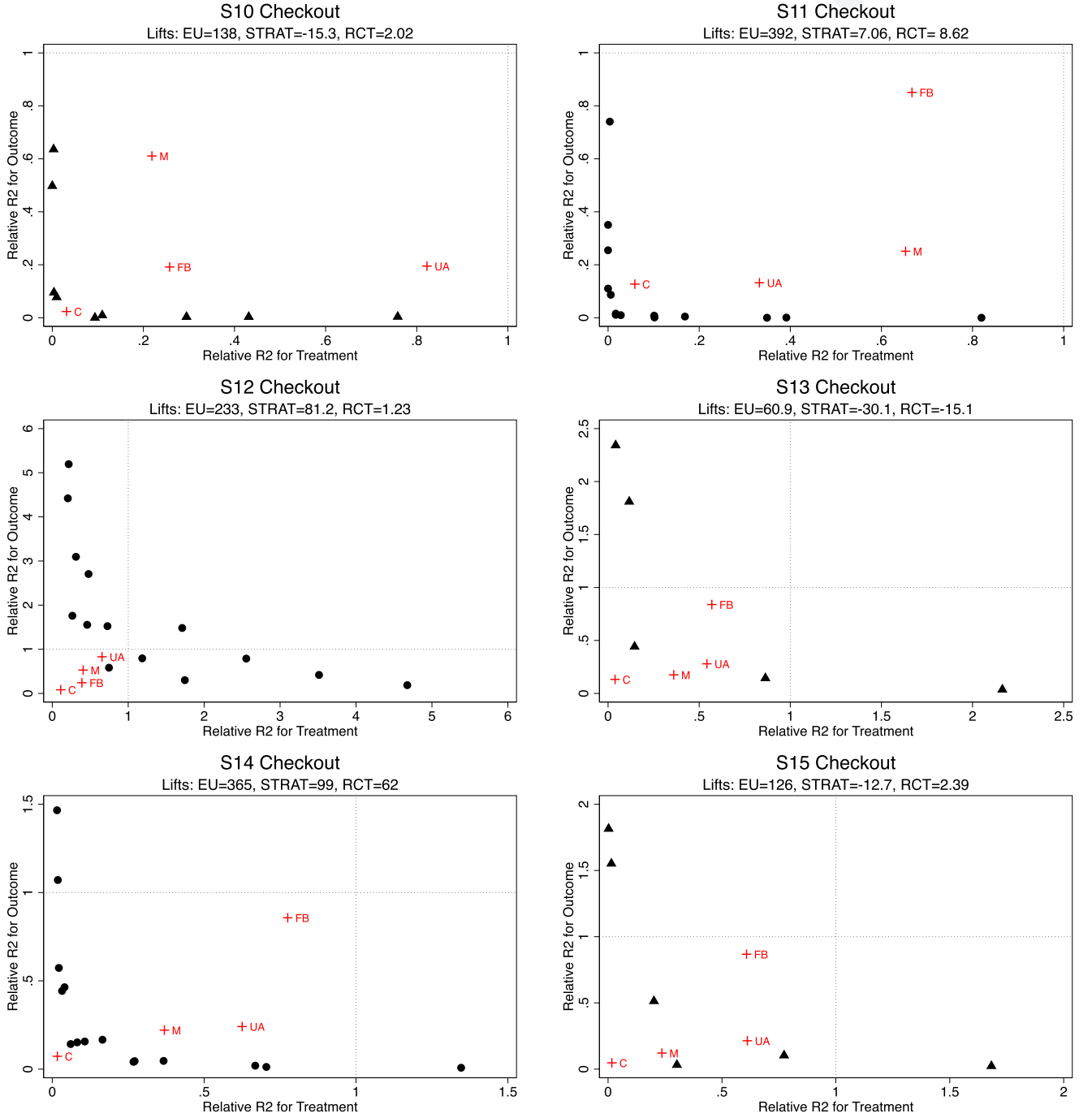
Appendix. Incorporating an Unobservable into an Observational Model

Our approach is based on the methodology developed in Rosenbaum and Rubin (1983a) and extended by Ichino et al. (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, $U \sim \text{Bern}(0.5)$, exists correlated with treatment and outcomes.²⁴ The observational model

Figure A.1. (Color online) Sensitivity Analysis for Studies 1–5 and 7–9

Q:31

Figure A.2. (Color online) Sensitivity Analysis for Studies 10–15



consists of a logit treatment equation and a linear outcome equation with a normal error term:

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

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

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:

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

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

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,

$$\hat{\tau}_{\text{rct}} = \tau(\alpha^*, \delta^*). \quad (\text{A.5})$$

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 et al. (2008), who propose directly specifying the parameters that characterize the distribution of the unobservable U :

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

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 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 et al. (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 $p_{01} > p_{00}$ because²⁵:

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

In the Rosenbaum–Rubin sensitivity model, α determines the strength of the unobservable to lead to selection into treatment. In Ichino et al. (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 $U_i = \{U_i^1, U_i^2, \dots, U_i^R\}$. Next, for each draw, we calculate an ATT over users and then average over the R ATT estimates. Specifically, for draw $r = 1, \dots, R$, we follow these steps:

- For each user, draw $U_i^r \sim \text{Bern}(p_{jk})$ 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, $\tau^{\text{sen},r}(\mathbf{p})$, as in Section 4.2.
- Calculate the average ATT over the draws, $\tau^{\text{sen}}(\mathbf{p}) = \frac{1}{R} \sum_r \tau^{\text{sen},r}(\mathbf{p})$.

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

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

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 re-define these terms using the notation in this appendix, such that the relative strength of the unobservable to affect outcomes is

$$R_{Y,\text{rel}}^2 = \frac{R_Y^2(\mathbf{p}) - R_Y^2(0)}{R_Y^2(0)}$$

and similar for $R_{W,\text{rel}}^2$.

Endnotes

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

² A growing literature focuses on measuring digital ad effectiveness using randomized experiments. See, for example, Goldfarb and Tucker (2011), Lewis and Reiley (2014), Sahni (2015), Sahni and Nair (2016), Johnson et al. (2016, 2017a, b, c), and Kalyanam et al. (2018). See Lewis et al. (2011) for a recent review.

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.

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 et al. 2010) and water-usage reduction (Ferraro and Miranda 2014, 2017).

⁵ Facebook refers to these ad tests as “conversion lift” tests (<https://www.facebook.com/business/a/conversion-lift>, accessed 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 April 7, 2018).

⁸ Additional factors beyond the advertiser’s bid determine the actual ranking. For more information, see <https://www.facebook.com/business/help/430291176997542>, accessed April 7, 2018.

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.

¹⁰ 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.

¹¹ 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, 2015) and machine learning (Athey, Imbens, and Wager, unpublished data). We leave to future work to explore how these methods perform in recovering experimental ad effects.

¹² 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 April 7, 2018.

¹³ If users in the test group shared ads with users in the control group, we are able to observe those impressions. Empirically, this rarely occurs. Any unobserved sharing would lead the RCT treatment effect estimates to be conservative because the shared ads could inflate the conversion rate in the control group.

¹⁴ See <https://www.facebook.com/business/help/164749007013531>, accessed April 7, 2018.

¹⁵ See <https://developers.facebook.com/docs/marketing-api/pacing>, accessed April 7, 2018.

¹⁶ More sophisticated specifications exist, including nonparametric models (Hirano et al. 2003), methods from machine learning (McCaffrey et al. 2004, Westreich et al. 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.

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 et al. 1997), are valid inferential techniques for matching estimators. Given this, we implement a subsampling procedure to estimate the ATT lift.

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” (p. 43).

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 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 et al. (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 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.

²³ For a related literature, see Murphy and Topel (1990), Altonji et al. (2005), and Oster (2016).

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.

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

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 $\tau^{cen}(\mathbf{p}^*) = \tau_{RCT}$. 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 $(1 - R_Y^2(0))$ and $(1 - R_W^2(0))$.

References

- Abadie A, Imbens GW (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 Bus. Rev.* 86(4):28.
- Altonji JG, Elder TE, Taber CR (2005) Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *J. Political Econom.* 113(1):151–184.
- Arceneaux K, Gerber AS, Green DP (2010) A cautionary note on the use of matching to estimate causal effects: An empirical example comparing matching estimates to an experimental benchmark. *Sociol. Methods Res.* 39(2):256–282.
- Athey S, Imbens GW, Wager S (2018) Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *J. Royal Statist. Soc. Ser. B* 80(4):597–623.
- Belloni A, Chen D, Chernozhukov V, Hansen C (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, Wood L (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, GotzeF, van ZweetW (1997) Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica* 7:1–31.

Q: 16

Q: 17

Q: 18

Q: 19

Q: 20

- Blake T, Nosko C, Tadelis S (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 BJ, Dubé J-P, Mela CF (2010) Do digital video recorders influence sales? *J. Marketing Res.* 47(6):998–1010.
- Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. *J. Econom. Surveys* 22(1):31–72.
- Cochran WG (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24(2): 295–314.
- comScore (2010) comScore announces introduction of AdEffx smart controlTM ground-breaking methodology for measuring digital advertising effectiveness. Press release.
- Cooper MJ, Gulen H, Rau PR (2005) Changing names with style: Mutual fund name changes and their effects on fund flows. *J. Finance* 60(6):2825–2858.
- Dehejia RH, Wahba S (2002) Propensity score matching methods for non-experimental causal studies. *Rev. Econom. Statist.* 84(1): 151–161.
- Draganska M, Hartmann WR, Stanglein G (2014) Internet versus television advertising: A brand-building comparison. *J. Marketing Res.* 51(5):578–590.
- Eckles D, Bakshy E (2017) Bias and high-dimensional adjustment in observational studies of Peer effects. Working paper, MIT Sloan School of Management, Cambridge, MA.
- Ellickson P, Collins RL, Hambarsoomians K, McCaffrey DF (2005) Does alcohol advertising promote adolescent drinking? Results from a longitudinal assessment *Addiction* 100(2):235–246.
- Ferraro PJ, Miranda JJ (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. *J. Econom. Behav. Organ.* 107:344–365.
- Ferraro PJ, Miranda JJ (2017) Panel data designs and estimators as substitutes for randomized controlled trials in the evaluation of public programs. *J. Assoc. Environ. Resource Econom.* 4(1): 281–317.
- Gelman A, Hill J (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 1st ed. (Cambridge University Press, Cambridge, UK).
- Gluck M (2011) Best practices for conducting online ad effectiveness research. Report, Interactive Advertising Bureau, New York.
- Goldfarb A, Greenstein S, Tucker C, eds. (2015) Measuring the effects of advertising: The digital frontier. *Economic Analysis of the Digital Economy* (University of Chicago Press, Chicago).
- Goldfarb A, Tucker C (2011) Online display advertising: Targeting and obtrusiveness. *Marketing Sci.* 30(3):389–404.
- Gu X, Rosenbaum P (1993) Comparison of multivariate matching methods: Structures, distances, and algorithms. *J. Comput. Graphical Statist.* 2(4):405–420.
- Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4):1161–1189.
- Ichino A, Fabrizia M, Nannicini T (2008) From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *J. Appl. Econom.* 23(3):305–327.
- Imai K, Ratkovic M (2014) Covariate balancing propensity score. *J. Royal Statist. Soc. Ser. B* 76(1):243–263.
- Imbens GW (2003) Sensitivity to exogeneity assumptions in program evaluation. *Amer. Econom. Rev.* 93(2):126–132.
- Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econom. Statist.* 86(1): 4–29.
- Imbens GW (2015) Matching methods in practice: Three examples. *J. Human Resources* 50(2):373–419.
- Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–475.
- Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1st ed. (Cambridge University Press, Cambridge, UK).
- Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J. Econom. Lit.* 47(1):5–86.
- Johnson GA, Lewis RA, Nubbemeyer EI (2017a) Ghost ads: Improving the economics of measuring ad effectiveness. *J. Marketing Res.* 54(6):867–884.
- Johnson GA, Lewis RA, Nubbemeyer E (2017b) The online display ad effectiveness funnel & carryover: Lessons from 432 field experiments. Working paper.
- Johnson GA, Lewis RA, Reiley D (2016) Location, location, location: Repetition and proximity increase advertising effectiveness. Working paper, University of Rochester, Rochester, NY.
- Johnson GA, Lewis RA, Reiley DH (2017c) When less is more: Data and power in advertising experiments. *Marketing Sci.* 36(1): 43–53.
- Kalyanam K, McAteer J, Marek J, Hodges J, Lin L (2018) Cross channel effects of search engine advertising on brick & mortar retail sales: Meta analysis of large scale field experiments on Google.com. *Quant. Marketing Econom.* 16(1):1–42.
- Klein C, Wood L (2013) Cross platform sales impact: Cracking the code on single source. Report, Nielsen Catalina Solutions and Time Inc, New York.
- Kumar A, Bezawada R, Rishika R, Janakiraman R, Kannan P (2016) From social to sale: The effects of firm-generated content in social media on customer behavior. *J. Marketing* 80(1):7–25.
- LaLonde RJ (1986) Evaluating the econometric evaluations of training programs with experimental data. *Amer. Econom. Rev.* 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, New York.
- Lewis R, Rao J (2015) The unfavorable economics of measuring the returns to advertising. *Quart. J. Econom.* 130(4):1941–1973.
- Lewis R, Reiley D (2014) Online ads and offline sales: Measuring the effect of retail advertising via a controlled experiment on Yahoo! *Quant. Marketing Econom.* 12(3):235–266.
- Lewis R, Rao J, Reiley D (2011) Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. *Proc. 20th Internat. Conf. World Wide Web* (Association for Computing Machines, New York), 157–166.
- McCaffrey DF, Ridgeway G, Morral AR (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* 9(4):403–425.
- Murphy K, Topel R (1990) Efficiency wages reconsidered: Theory and evidence. Weiss Y, Fishelson G, eds. *Advances in the Theory and Measurement of Unemployment*, Vol. 2 (Palgrave Macmillan, London), 204–240.
- Oster E (2016) Unobservable selection and coefficient stability: Theory and evidence. *J. Bus. Econom. Statist.* 1–18.
- Pham TT, Shen Y (2017) A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform. Working paper, Stanford GSB, <https://arxiv.org/pdf/1706.02795.pdf>.
- Politis DN, Romano JP (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* 22(4): 2031–2050.
- Robins J, Ritov Y (1997) Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statist. Medicine* 16(3):285–319.
- Rosenbaum P, Rubin DB (1983a) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Royal Statist. Soc. Ser. B* 45(2):212–218.

Q: 22, 23

Q: 24

- Rosenbaum PR, Rubin DB (1983b) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.
- Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* 39(1):33–38.
- Rubin DB (1978) Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* 6(1):34–58.
- Rubin DB (2001) Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services Outcomes Res. Methodol.* 2:169–188.
- Rubin DB, Waterman RP (2007) Estimating the causal effects of marketing interventions using propensity score methodology. *Statist. Sci.* 21(2):206–222.
- Sahni N (2015) Effect of temporal spacing between advertising exposures: Evidence from online field experiments. *Quant. Marketing Econom.* 13(3):203–247.
- Sahni N, Nair H (2016) Native advertising, sponsorship disclosure and consumer deception: Evidence from mobile search-ad experiments Working paper, Stanford GSB <http://ssrn.com/abstract=2737035>.
- Stuart EA (2010) Matching methods for causal inference: A review and a look forward. *Statist. Sci.* 25(1):1–21.
- Stuart EA, Rubin DB (2007) Matching methods for causal inference: Designing observational studieschap. *Best Practices in Quantitative Methods* (Sage Publishing, Thousand Oaks, CA).
- Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J. Royal Statist. Soc.* 58(1): 267–288.
- Westreich D, Lessler J, Funk MJ (2010) Propensity score estimation: Machine learning and classification methods as alternatives to logistic regression. *J. Clin. Epidemiol.* 63(8):826–833.
- Wooldridge JM (2007) Inverse probability weighted estimation for general missing data problems. *J. Econom.* 141(2)1281–1301.
- Zubizarreta JR (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* 107(500):1360–1371.
- Zubizarreta JR (2015) Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* 110(511):910–922.

AUTHOR QUERIES

DATE 1/4/2019

JOB NAME MKSC

ARTICLE 20181135

QUERIES FOR AUTHOR Gordon et al.

THIS QUERY FORM MUST BE RETURNED WITH ALL PROOFS FOR CORRECTIONS

- Q: 1_Please provide a "short title" for your article, of no more than 60 characters, including spaces.
- Q: 2_Please provide all funding information, if applicable, including the names of the institutions as well as any grant numbers associated with the funding, or confirm that no funding was received.
- Q: 3_Color figures will be converted to grayscale for the print version unless color in print has been requested and paid for. Authors have the responsibility to check that the figures are sufficiently clear in both the online color and print black and white versions.
- Q: 4_Please confirm that the article title, author names, affiliations, and email addresses are set correctly. If applicable, please provide author ORCID numbers.
- Q: 5_Please confirm that keywords are correct as set.
- Q: 6_Please confirm that heading levels are correct as set.
- Q: 7_Please verify that all displayed equations and in-text math notations are set correctly.
- Q: 8_Please spell out GMM.
- Q: 9_Please spell out LASSO at first use in text, with abbreviation following in parentheses.
- Q: 10_Regarding "orange circles" in description of Figure 6: because the figures will be printed in grayscale, please revise text to avoid use of color descriptors.
- Q: 11_Regarding text "which is highlighted in red": see previous query regarding use of color descriptors for figures to be printed in grayscale. Check entire paragraph.
- Q: 12_Referring to Appendix items by their manuscript page number will not make sense: page numbers will change when your paper appears in print. Generic "can be found in the Appendix" has been substituted here. OK? See subsequent queries.
- Q: 13_Is "in Figure A.1" correct here, or is more than one figure being cited? If the latter, please revise for accuracy.
- Q: 14_Please spell out organization acronyms in the Acknowledgment, unless those are the official names.
- Q: 15_Is "(5)" as intended here, or is "(A.5)" meant? Please check all equation citations in the Appendix and confirm accuracy.

- Q: 16_Please provide last page number for reference "Abraham 2008."
- Q: 17_The references "Athey et al. 2018" and "Goldfarb et al. 2015" were not cited in the text. We have retained each as provided by the author. Please add a citation or delete the references.
- Q: 18_Please provide complete information for reference "Benady 2016" and "Bond 2017": Month and day of publication, and URL (web address) or page number.
- Q: 19_Please provide accessed date, publisher and its location for reference "Berkovich and Wood 2016."
- Q: 20_Please provide the issue number for each of the references "Bickel et al. 1997," "Ferraro and Miranda 2014," "Oster 2016," and "Rubin 2001."
- Q: 21_For reference "comScore 2010" please provide the date of the press release and the URL, if available.
- Q: 22_Please update the working papers to reflect publication if possible.
- Q: 23_Please provide publisher and its location for reference "Johnson et al., 2017b".
- Q: 24_Please provide publisher location and accessed date for references "Pham and Shen 2017" and "Sahni and Nair 2016."
- Q: 25_Your color figures appear in color in your page proof to simulate their presentation in color online, but they will be converted to grayscale in the print version of your article because you have not requested color processing for print. Please (a) check your figures to confirm that they will be sufficiently clear in grayscale presentation in the print version of your article and (b) modify your figure legends as necessary to remove any references to specific colors in the figure.
- Q: 26_Please check that all tables and figures, including their titles and notes (which may contain edits) are set correctly.
- Q: 27_Please confirm that you have requested and received permission to adapt/reprint Figures 1 and 2.
- Q: 28_Legend to Figure 6 ("The sets of variables correspond to those described on page 21."): Please replace manuscript page number, which won't correspond to printed article, with a specific section number or table.
- Q: 29_Legend to Figure 8: please check and confirm accuracy. Is symbol missing from empty square brackets? If not, please replace [] with "No symbol means we fail to reject the hypothesis" and move to end of legend. Same query applies to Figure 9.
- Q: 30_Figure 10: Please revise figure "key" so that it does not use color descriptors (image will be printed in grayscale).
- Q: 31_Figure A.2 is not cited in the text. Please add a citation for this figure.
- Q: 32_If possible, please provide a version of Figure 10 with larger font.