Volume Title: Economic Analysis of the Digital Economy

Volume Author/Editor: Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-20684-X; 978-0-226-20684-4

Volume URL: http://www.nber.org/books/gree13-1

Conference Date: June 6–7, 2013

Publication Date: April 2015

Chapter Title: Measuring the Effects of Advertising: The Digital Frontier

Chapter Author(s): Randall Lewis, Justin M. Rao, David H. Reiley

Chapter URL: http://www.nber.org/chapters/c12991

Chapter pages in book: (p. 191 – 218)

7

Measuring the Effects of Advertising
The Digital Frontier

Randall Lewis, Justin M. Rao, and David H. Reiley

7.1 Introduction

In the United States, advertising is a $200 billion industry, annually. We all consume "free" services—those monetized by consumer attention to advertising—such as network television, e-mail, social networking, and a vast array of online content. Yet despite representing a relatively stable 2 percent of gross domestic product (GDP) since World War I and subsidizing activities that comprise most of Americans' leisure time (Bureau of Labor Statistics 2010), advertising remains poorly understood by economists. This is primarily because offline data have typically been insufficient for a firm (or researcher) to measure the true impact of advertising on consumer purchasing behavior. Theories of advertising (Demsetz 1982; Kessides 1986; Becker and Murphy 1993) that have important implications for competition are even harder to empirically validate. The digital era offers an unprecedented opportunity to bridge this informational divide. These advances, both realized and potential, can be attributed to two key factors: (1) individual-level data on ad delivery and subsequent purchasing behavior can be linked and made available to advertisers at low cost; and (2) ad delivery can be randomized at the individual level, generating exogenous variation essential to

Randall Lewis is an economic research scientist at Google, Inc. Justin M. Rao is an economic researcher at Microsoft Research. David H. Reiley is a research scientist at Google, Inc.

Much of this work was done when all the authors were at Yahoo! Research. We thank Garrett Johnson, Dan Nguyen, Sergiy Matusevych, Iwan Sakran, Taylor Schreiner, Valter Sciarillo, Christine Turner, Michael Schwarz, Preston McAfee, and numerous other colleagues for their assistance and support in carrying out the research. For acknowledgments, sources of research support, and disclosure of the authors' material financial relationships, if any, please see http://www.nber.org/chapters/c12991.ack.

identifying causal effects.[1] In this chapter we explore the dramatic improvement in the empirical measurements of the returns to advertising, highlight fundamental challenges that currently remain, and look to what solutions we think the future will bring.

Digital advertising has led to standard reporting of precise quantitative data for advertising campaigns, most notably the click-through rate (CTR). Of course, the CTR of an ad is only an intermediate proxy for the real outcome of interest to the advertiser: increased purchases by consumers, both in the present and future.[2] Despite these limitations, intermediate metrics such as the CTR have proved to be enormously useful dependent variables in automated targeting algorithms that match ads with consumers and contexts (Pandey and Olston 2006; Gonen and Pavlov 2007). Related intermediate metrics come from "purchasing intent" surveys paired with randomized exposure to a firm's advertising. Cross-experiment analysis of such surveys has provided estimates of the relative value of targeted (versus untargeted) advertising (Goldfarb and Tucker 2011b), contextual relevance and ad intrusiveness (Goldfarb and Tucker 2011a), and has informed the debate on privacy (Tucker 2012).

The advances in both academic understanding and business best-practice attributable to these intermediate metrics should not be understated. But while general insights on how ad features impact users can guide advertising spend and CTR maximizing algorithms can make spending more efficient, a firm is presumably interested in measuring the overall returns on advertising investment: dollars of sales causally linked to the campaign versus dollars spent. An overreliance on intermediate metrics can draw attention away from the true underlying goal, and research has shown it can lead to highly suboptimal spending decisions (Blake, Nosko, and Tadelis 2014).

Along with deficiencies in intermediate metrics, endogeneity of advertising exposure is the other key challenge in measuring advertising returns. Traditional econometric measurements typically rely on aggregate data fraught with identification problems due to the targeted nature of advertising (Bagwell 2007).[3] Yet despite the ability to run very large randomized control trials made possible by digital delivery and measurement, we have discovered a number of conceptual flaws in standard industry data collection and anal-

1. There have been experimental approaches to measuring advertising effectiveness in the past, see most notably the split-cable experiments of Lodish et al. (1995), but these were typically conducted as small pilots and not using the normal ad delivery pipeline.

2. Toward these ends, advertisers use browser cookies and click beacons to obtain a "conversion rate," the ratio of transactions attributed to the campaign to ad exposures. This measure seems ideal, but the attribution step is critical and current methods of assigning attribution have serious flaws, which we discuss in detail.

3. The split cable TV experiments reported in Lodish et al. (1995) are a notable exception. The sample sizes in these experiments, run in a small US town, were far smaller than online experiments, and the authors did not report per experiment confidence intervals, rather they used cross-experiment techniques to understand what factors tended to influence consumers (for a follow-up analysis, see Hu, Lodish, and Krieger [2007]).

ysis methods used to measure the effects of advertising. In other words, the deluge of data on advertising exposures, clicks, and other associated outcomes have not necessarily created greater understanding of the basic causal effects of advertising, much less an understanding of more subtle questions such as the relative effectiveness of different types of consumer targeting, ad creatives, cross-channel effects, or frequency of exposure. The voluminous data, it seems to us, have not only created opportunity for intelligent algorithmic advances, but also mistaken inference under the guise of "big data."

First, many models assume that if you do not click on the ad, then the ad has no effect on your behavior. Here we discuss work by coauthors Lewis and Reiley that showed online ads can drive offline sales, which are typically not measured in conversion or click rates; omitting these nonclick-based sales leads to underestimating the total effects of advertising. Linking online and offline sales requires a dedicated experimental infrastructure and third-party data merging that have only recently become possible.

Second, many models assume that if you do click on an ad and subsequently purchase, that conversion must have been *due to that ad*. This assumption seems particularly suspect in cases, such as search advertising, where the advertising is deliberately targeted at those consumers most likely to purchase the advertised product and temporally targeted to arrive when a consumer is performing a task related to the advertised good. Research has shown, for example, that a person searching for "ebay shoes" is very likely to purchase shoes on eBay regardless of the intensity of advertising (Blake, Nosko, and Tadelis 2014). While this is an extreme example, Blake, Nosko, and Tadelis (2014) also show that the problem arises generally, and measuring the degree to which advertising crowds out "organic conversions" is difficult to measure precisely. Näive approaches effectively assume this problem away, but since only "marginal clicks" are valuable and all clicks count toward the CTR, these methods will always overstate the causal effect on users who clicked the ad.

Third, more sophisticated models that do compare exposed to unexposed users to establish a baseline purchase rate typically rely on natural, endogenous advertising exposure and can easily generate biased estimates due to unobserved heterogeneity (Lewis, Rao, and Reiley 2011). This occurs when the pseudo-control group does not capture important characteristics of the treated group, such as purchase intent or browsing intensity, which we show can easily be correlated with purchases whether advertising is present or not. Using data from twenty-five large experiments run at Yahoo! (Lewis and Rao 2013), we have found that the standard deviation of purchases is typically ten times the mean. With such a noisy dependent variable, even a tiny amount of endogeneity can severely bias estimates. Beyond inducing bias in coefficient estimates, these specification errors also give rise to an overprecision problem. Because advertising typically explains only a very small fraction of the variance in consumer transaction behavior, even cleanly designed experiments typically require over a million subjects in order to be able to measure economically meaningful effects with any statistical precision (but even experiments with one million subjects can have surprisingly weak power, depending on the variance in sales).

Since experiments are generally considered the gold standard for precision[4] (treatment is exogenous and independent across individuals), we should be suspicious if observational methods claim to offer higher precision. Further, with nonexperimental methods, omitted heterogeneity or selection bias (so long as it can generate a partial R-squared of 0.00005 or greater) can induce bias that swamps plausible estimates of advertising effectiveness. Thus, if an advertiser does not use an experiment to evaluate advertising effectiveness, she has to have a level of confidence in her model that, frankly speaking, we find unreasonable given the obvious selection effects due to ad targeting and synchronization of advertising with product launches (e.g., new iPad release) and demand shocks (e.g., holiday shopping season).

Experimental work on measuring the dollar returns to advertising has given us a deeper appreciation for the limits of current data and methods. For example, we show that seemingly simple "cross-channel" complementarity measures are exceedingly difficult to reliably estimate. Here we present evidence taken from Lewis and Nguyen (2013) that display advertising can increase keyword searches for the advertised brand. Some clicks on sponsored links are incorrectly attributed entirely to the search ad, but while the directional impact on searches can be documented, we cannot tell if search ads perform better or worse in terms of the conversion rate when paired with display advertising. A similar experimental design at a much larger scale could answer this sort of question, but advertising to over five to ten million individuals may be out of reach[5] for most advertisers. These findings are confirmed by similar work on online advertising spillovers (Rutz and Bucklin 2011; Papadimitriou et al. 2011).

So while some questions are answerable with feasible (at least for some market participants) scale, we believe other questions are still outside the statistical power of current experimental infrastructure and methods. The most prominent example is the long-run effects of advertising. Essentially any analysis of the impact of advertising has to make a judgment call on which time periods to use in the analysis. Often this is the "campaign window" or the campaign window plus a chosen interval of time (typically one to four weeks). These thresholds are almost certainly "wrong" because any impact that occurs after the cutoff should count in the return on investment (ROI) calculation. We explain why practitioners typically choose relatively short impact windows. The intuition is that the longer the time window

4. Not all experiments are created equal and methodologies to use preexperiment data to enhance power as well as postexperiment trimming have advanced considerably in the digital era (Deng, Kohavi, and Walker 2013).

5. Pun intended.

under study, the lower the signal-to-noise ratio in the data (presuming the ad gets less impactful over time): point estimates of the cumulative effect tend to increase with longer time horizons, but standard errors of the effect increase by even more. This leads to an estimation "impossibility" analogous to the well-known "curse of dimensionality."

In the next two sections we shift our gaze further into the future. First, we discuss how computational methods have increased advertising effectiveness through automated targeting and bidding. With automated targeting, the conversation is usefully shifted from "who to hit" to "what should I get." Currently, the key parameters of the automated system such as the valuation of actions such as clicks or conversions, the budget of the campaign and the duration, must still be entered by a human. Indeed, these are the exact parameters that we have argued are very difficult to estimate. However, there is no major technical barrier to incorporating controlled randomization—on the fly experimentation—into the core algorithm. By constantly incorporating experimentation, an informative prior could be developed and returns could be more precisely estimated (which would then govern bid, budget, and so forth). To unlock the full potential of this class of algorithms, ad exchanges would have to provide data to participants on the outcomes of auctions in which the bidder intentionally lost. Currently, outcome tracking is only possible if you win the auction, meaning today this type of experimentation is limited to temporal and geography-based identification, severely limiting power. In our final section we extend the discussion on how advances in ad delivery, measurement, and infrastructure are creating opportunities to advance the science of advertising. We discuss how the provision of these features and data relates to the incentives facing the advertising platform. In the final section we present concluding remarks.

7.2 Selection and Power

In today's dollars, the average American is exposed to about $500 worth of advertising per year.[6] To break even, the universe of advertisers needs to net about $1.35 in *marginal profits* per person per day. Given the gross margins of firms that advertise, our educated guess is that this roughly corresponds to about four to six dollars in incremental sales per day.

When an advertiser enters this fray, it must compete for consumers' attention. The cost per person of a typical campaign is quite low. Online "display" (banners, rectangular units, etc.) campaigns that deliver a few ads per day to a targeted individual cost about one to two cents per person per day. Televi-

6. Mean GDP per American is approximately $50,000 in 2011, but median household income is also approximately $50,000. The average household size is approximately 2.5, implying an individual's share of median household income is roughly $20,000. Thus, while 2 percent of GDP actually implies a per capita expenditure of $1,000, we use $500 as a round and conservative figure that is more representative of the average American's ad exposure.

sion ads delivered once per person per day are only a bit more expensive. Note that even an aggressive campaign will typically only garner a small percentage of an individual's daily advertising exposure. We see many ads per day and presumably only a minority of them are relevant enough to a given person to impact his behavior.

The relatively modest average impact per person makes it difficult to assess costeffectiveness. What complicates matters further is that individual-level sales are quite volatile for many advertisers. An extreme example is automobiles—the sales impact is either tens of thousands of dollars or zero.[7] While not as extreme, many other heavily advertised categories, including consumer electronics, clothing and apparel, jewelry, air travel, banking, and financial planning also have volatile consumption patterns.[8] Exceptions to this class are single goods sold through direct conversion channels. Here we summarize work presented in Lewis and Rao (2013), which used twenty-five large advertising field experiments to quantify how individual expenditure volatility impacts the power of advertising effectiveness (hereafter, adfx) experiments. In general, the signal-to-noise ratio is much lower than we typically encounter in economics.

We now introduce some formal notation to clarify the argument. Consider an outcome variable y (sales), an indicator variable x equal to 1 if the person was exposed to the advertising, and a regression estimate, $\hat{\beta}$, of the average difference between the exposed (E) and unexposed (U) groups. In an experiment, exposure is exogenous—determined by a flip of the proverbial coin. In an observational study, one would also condition on covariates W, which could include individual fixed effects, and the following notation would use $y \mid W$. All the following results go through with the usual "conditional upon" caveat. We consider a regression of y on x, whose coefficient $\hat{\beta}$ will give us a measure of the average dollar impact of the advertising per consumer.

We use standard notation for the sample means and variances of the sales of the exposed and unexposed groups, the difference in means between those groups, and the estimated standard error of that difference in means. We assume for simplicity that the exposed and unexposed samples are the same size ($N_E = N_U = N$) as well as equal variances ($\sigma_E = \sigma_U = \sigma$) to simplify the formulas:

$$\bar{y}_E \equiv \frac{1}{N_E}\sum_{i \in E} y_i,\ \bar{y}_U \equiv \frac{1}{N_U}\sum_{i \in U} y_i \tag{1}$$

$$\hat{\sigma}_E^2 \equiv \frac{1}{N_E - 1}\sum_{i \in E}(y_i - \bar{y}_E)^2,\ \hat{\sigma}_U^2 \equiv \frac{1}{N_U - 1}\sum_{i \in U}(y_i - \bar{y}_U)^2 \tag{2}$$

7. The marginal profit impact is large, but clearly smaller, as it is the gross margin times the sales impact.

8. For a bank, the consumption pattern once you sign up might be predictable, but the bank is making money from consumer switching, which is "all or nothing."

$$\Delta\bar{y} \equiv \bar{y}_E = \bar{y}_U \tag{3}$$

$$\hat{\sigma}_{\Delta\bar{y}} \equiv \sqrt{\frac{\hat{\sigma}_E^2}{N_E} + \frac{\hat{\sigma}_U^2}{N_U}} = \sqrt{\frac{2}{N}} \cdot \hat{\sigma}. \tag{4}$$

We focus on two familiar econometric statistics. The first is the R^2 of the regression of y on x, which gives the fraction of the variance in sales explained by the advertising (or, in the model with covariates, the partial R^2 after first partialing out covariates—for more explanation, see Lovell [2008]):

$$R^2 = \frac{\sum_{i\in U}(\bar{y}_U - \bar{y})^2 + \sum_{i\in E}(\bar{y}_E - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{2N[(1/2)\Delta\bar{y}]^2}{2N\hat{\sigma}^2} = \frac{1}{4}\left(\frac{\Delta\bar{y}}{\hat{\sigma}}\right)^2. \tag{5}$$

Second is the t-statistic for testing the hypothesis that the advertising had no impact:

$$t_{\Delta\bar{y}} = \frac{\Delta\bar{y}}{\hat{\sigma}_{\Delta\bar{y}}} = \sqrt{\frac{N}{2}}\left(\frac{\Delta\bar{y}}{\hat{\sigma}}\right) \tag{6}$$

In both cases, we have related a standard regression statistic to the ratio between the average impact on sales and the standard deviation of sales between consumers.

In the following hypothetical example, we calibrate values using approximately median values from nineteen retail sales experiments run at Yahoo!. For expositional ease, we will discuss it as if it is a single experiment. The campaign goal is a 5 percent increase in sales during the two weeks of the campaign, which we will use as our "impact period" of interest. During this period, customers of this advertiser make purchases with a mean of \$7 and a standard deviation of \$75.[9] The campaign costs \$0.14 per customer, which amounts to delivering 20–100 display ads at a price of \$1–\$5 CPM,[10] and the gross margin (markup over cost of goods sold, as a fraction of price) is assumed to be about 50 percent.[11] A 5 percent increase in sales equals \$0.35 per person, netting profits of \$0.175 per person. Hence, the goal for this campaign is to deliver a 25 percent return on investment (ROI): \$0.175/\$0.14 = 1.25.[12]

The estimation challenge facing the advertiser in this example is to detect a \$0.35 difference in sales between the treatment and control groups amid

9. Based on data-sharing arrangements between Yahoo! and a number of advertisers spanning the range from discount to high-end retailers, the standard deviation of sales is typically about ten times the mean. Customers purchase goods relatively infrequently, but when they do, the purchases tend to be quite large relative to the mean.

10. CPM is the standard for impression-based pricing for online display advertising. It stands for "cost per mille" or "cost per thousand"; M is the Roman numeral for 1,000.

11. We base this assumption on our conversations with retailers and our knowledge of the industry.

12. For calibration purposes, note that if the gross margin were 40 percent instead of 50 percent, this would imply a 0 percent ROI.

the noise of a \$75 standard deviation in sales. The ratio is very low: 0.0047. From our derivation above, this implies an R^2 of:

$$R^2 = \frac{1}{4} \cdot \left(\frac{\$0.35}{\$75}\right)^2 = 0.0000054. \tag{7}$$

That is, even for a *successful* campaign with a *relatively large* ROI, we expect an R^2 of only 0.0000054. This will require a very large N to identify any influence at all of the advertising, let alone give a precise confidence interval. Suppose we had two million unique users evenly split between test and control in a fully randomized experiment. With a true ROI of 25 percent and a ratio of 0.0047 between impact size and standard deviation of sales, the expected t-stat is 3.30, using the above formula. This corresponds to a test with power of about 95 percent at the 10 percent (5 percent one-sided) significance level, as the normally distributed t-statistic should be less than the critical value of 1.65 about 5 percent of the time given the true effect is a 25 percent ROI. With 200,000 unique customers, the expected t-statistic is 1.04, indicating the test is hopelessly underpowered to reliably detect an economically relevant impact: under the alternative hypothesis of a healthy 25 percent ROI, we fail to reject the null 74 percent of the time.[13]

The low $R^2 = 0.0000054$ for the treatment variable x in our hypothetical randomized trial has serious implications for observational studies, such as regression with controls, difference-in-differences, and propensity score matching. A very small amount of endogeneity would *severely bias* estimates of advertising effectiveness. An omitted variable, misspecified functional form, or slight amount of correlation between browsing behavior and sales behavior generating R^2 on the order of 0.0001 is a *full order of magnitude* larger than the true treatment effect. Compare this to a classic economic example such as the Mincer wage/schooling regression (Mincer 1962), in which the endogeneity is roughly 1/8 the treatment effect (Card 1999). For observational studies, it is always important to ask, "What is the partial R^2 of the treatment variable?" If it is very small, as in the case of advertising effectiveness, clean identification becomes paramount, as a small amount of bias can easily translate into an economically large impact on the coefficient estimates.

Our view has not yet been widely adopted, however, as evidenced by the following quotation from the president of comScore, a large data provider for online advertising:

> Measuring the online sales impact of an online ad or a paid-search campaign—in which a company pays to have its link appear at the top of a page of search results—is straightforward: We determine who has viewed

13. Note that when a low-powered test does, in fact, correctly reject the null, the point estimates conditional on rejecting will be significantly larger than the alternatively hypothesized ROI. See Gelman and Carlin (2013) regarding this "exaggeration factor."

the ad, then compare online purchases made by those who have and those who have not seen it. (Abraham 2008)

The argument we have made shows that simply comparing exposed to unexposed can lead to bias that is many orders of magnitude larger than the true size of the effect. Indeed, this methodology led the author to report as much as a 300 percent improvement in outcomes for the exposed group, which seems surprisingly high (it would imply, for instance, that advertisers are grossly underadvertising). Since all ads have some form of targeting,[14] endogeneity is always a concern. For example, most display advertising aims to reach people likely to be interested in the advertised product, where such interest is inferred using demographics or past online behavior of that consumer. Similarly, search advertising targets consumers who express interest in a good at a particular point in time, where the interest is inferred from their search query (and potentially past browsing behavior). In these cases, comparing exposed to unexposed is precisely the *wrong* thing to do. By creating exogenous exposure, the first generation of advertising experiments have been a step in the right direction. Experiments are ideal—necessary, in fact—for solid identification.

Unfortunately, for many advertised products the volatility of sales means that even experiments with millions of unique users can still be underpowered to answer basic questions such as "Can we reject the null hypothesis that the campaign had zero influence on consumers' purchasing behavior?" Measuring sales impact, even in the short run, turns out to be much more difficult than one might have thought. The ability to randomize ad delivery on an individual level and link it to data on customer-level purchasing behavior has opened up new doors in measuring advertising effectiveness, but the task is still by no means easy. In the remainder of the chapter we discuss these challenges. The next section focuses on using the right metrics to evaluate advertising.

7.3 The Evolution of Advertising Metrics

The click-through-rate, or CTR, has become ubiquitous in the analysis and decision making surrounding online advertising. It is easy to understand why: clicks are cleanly defined, easily measurable, and occur relatively frequently. An obvious but intuitively appealing characteristic is that an ad click cannot occur in the absence of an ad. If one runs 100,000 ads and gets a 0.2 percent CTR (a typical rate for a display ad or a low-ranked search ad), it is tempting to conclude the ad caused 200 new website visits. The assump-

14. "Untargeted" advertising usually has implicit audience targeting based on where the ads are shown or implicit complementary targeting due to other advertisers purchasing targeted inventory and leaving the remnant inventory to be claimed by advertisers purchasing "untargeted" advertising inventory.

tion may well be true for new or little-known brands. But for well-known advertisers, there are important ways that consumers might navigate to the site in the absence of an ad, such as browsing directly to the site by typing the name in the URL window of the browser or finding it in organic (that is, not paid or "sponsored") search results on a topic like "car rental." It is a mistake to assume that *all* of those 200 visits would not have occurred in the absence of the ad—that is, those clicks may be crowding out visits that would have happened via other means (Kumar and Yildiz 2011; Chan et al. 2010).

The overcounting problem is surmountable with randomized trials where the control group is used to estimate the "baseline arrival rate." For example, a sponsored search ad could be turned off during random times of the day and the firm could measure arrivals from the search engine for when the ad is running and when it is not (this approach is used in Blake, Nosko, and Tadelis [2014]).[15] A deeper problem with the CTR is what it misses. First, it does little for "brand advertisers"—firms that are not trying to generate immediate online sales, but rather to promote awareness and goodwill for the brand. To assess their spend, brand advertisers have traditionally relied on surveys that attempt to measure whether a campaign raised the opinion of the firm in the minds of their target consumers (Goldfarb and Tucker 2011b). Linking the surveys to future purchasing behavior adds another layer of complexity, both because the time frame from exposure to sale is longer (something we will discuss in more detail in section 7.5) and because it requires a reliable link from hypothetical responses to actual behavior, which can be fraught with what is known as "hypothetical bias" (Dickie, Fisher, and Gerking 1987; Murphy et al. 2005). One common approach to neutralize hypothetical bias is to use the surveys to make relative comparisons between campaigns.

For advertisers that sell goods both online and in brick-and-mortar stores the click (or online conversions) can be a poor proxy for overall ROI. Lewis and Reiley (2013a) show that for a major retailer, the *majority* of the sales impact comes offline. Johnson, Lewis, and Reiley (2013) link the offline impact to consumers who lived in close physical proximity to one of the retailer's locations. These studies indicate purely online measurements can induce a large negative bias in measuring the returns to advertising. For firms that do business on- and offline it is essential to develop the infrastructure to link online ad exposure to offline sales.

An alternative to the click is the further downstream outcome measure known as a "customer acquisition" (which itself might be considered a short-term proxy for the net-present-discounted value of a customer). Advertisers can now run "cost per acquisition" (CPA) advertising on many

15. Despite the simplicity of their design, Blake, Nosko, and Tadelis (2014) estimate that their employer, eBay, had been wasting tens of millions of dollars a year.

ad exchanges.[16] An acquisition, or conversion, is defined as a successful transaction that has a "qualifying connection" to the advertisement. On the surface, focusing on conversions seems more attractive than clicks because it is a step closer to sales. Unfortunately, this benefit brings with it what is known as the "attribution problem": which ad gets "credit" for a given sale? Suppose a consumer views and clicks a given ad, but does not purchase on the same day. Over the next few days, she sees a host of other ads for the product (which is likely, given a practice known as "retargeting") and then purchases the good. Which ad should get credit for the purchase?

Ad exchanges tend to use a set of rules to solve these problems from an accounting perspective. Common rules include requiring a click for credit or only counting the "last click" (so if a consumer clicks a retargeted ad, that ad gets credit). Requiring a click seems to make sense and is enormously practical as it means a record of all viewers that see the ad but do not click need not be saved.[17] However, requiring a click errs in assuming that ads can only have an impact through clicks, which is empirically not true (Lewis, Reiley, and Schreiner 2012). The "last click" rule also has intuitive appeal. The reasoning goes as follows: had the last click not occurred, the sale would not have happened. Even if this were true, which we doubt, the first click or ad view might have led to web search or other activity, including the behavioral markers used for retargeting, which made the last click possible. The causal attribution problem is typically solved by ad hoc rules set by the ad exchange or publisher such as "the first ad and the last ad viewed before purchase each get 40 percent of the credit, while the intermediate ad views share the remaining 20 percent of the credit for the purchase."[18] A proliferation of such rules gives practitioners lots of choices, but none of them necessarily gives an unbiased measurement of the performance of their ad spending. In the end, such complicated payment rules might make the click more attractive after all.

The attribution problem is also present in the question of complementaries between display and search advertising. Recent work has shown that display ads causally influence search behavior (Lewis and Nguyen 2013). The authors demonstrate this by comparing the search behavior of users exposed to the campaign ad to users who would have been served the campaign ad but were randomly served a placebo. Brand-related keywords were significantly more prevalent in the treatment group as compared to the control. The attribution problem has received more attention in online advertising because of the popularity of cost-per-acquisition and cost-per-click payment mechanisms, but it applies to offline settings as well. How do we

16. But not the major search engines, as of August 2013.

17. A CTR of ≈ 0.2 percent meaning, storage, and processing costs of only clicks involves only 1/500 of the total ad exposure logs.

18. Source: https://support.google.com/analytics/bin/answer.py?hl=en\&answer=1665189.

know, for example, whether an online ad was more responsible for an online conversion than was the television ad that same user saw? Nearly every online campaign occurs contemporaneously with a firm's offline advertising through media such as billboards and television because large advertisers are continuously advertising across many media.[19] Directly modeling the full matrix of first-order interactions is well beyond the current state of the art. Indeed in every paper we know of evaluating online advertising, the interactions with offline spending is ignored.

Our discussion thus far has indicated that the evolution of advertising metrics has brought forth new challenges linking these metrics to the causal impact on sales. However, one way in which intermediate metrics have proved unambiguously useful for advertisers is providing relatively quick feedback on targeting strategies allowing for algorithmic adjustments to the ad-serving plan. For instance, while it may be unreasonable to assume that the click captures all relevant effects of the ad, it may very natural to assume that within a given class of advertisements run by a firm a higher CTR is always preferred to a lower one. If so, bandit algorithms can be applied to improve the efficiency of advertising spend and give relative comparisons of campaign effectiveness, allowing one to prioritize better performing advertisements (Pandey and Olston 2006; Gonen and Pavlov 2007). We discuss these advances in more detail in section 7.7.

7.4 A Case Study of a Large-Scale Advertising Experiment

To get a better idea of how large advertising experiments are actually run, in this section we present a case study taken from Lewis and Reiley (2013a) (herein "LR"). Lewis and Reiley ran a large-scale experiment for a major North American retailer. The advance the paper makes is linking existing customers in the retailer's sales records, for both online and brick-and-mortar sales, to a unique online user identifier, in this case the customer's Yahoo! username.

The experiment was conducted as follows. The match yielded a sample of 1,577,256 individuals who matched on name and either e-mail or postal address. The campaign was targeted only to existing customers of the retailers as determined by the match. Of these matched users, LR assigned 81 percent to a treatment group who subsequently viewed two advertising campaigns promoting the retailer when logged into Yahoo's services. The remaining 19 percent were assigned to the control group and prevented from seeing any of the retailer's ads from this campaign on the Yahoo! network of sites. The simple randomization was designed to make the treatment-control assignment independent of all other relevant variables.

19. Lewis and Reiley (2013b) show that Super Bowl commercials cause viewers to search for brand-related content across a wide spectrum of advertisers.

Table 7.1 **Summary statistics for the campaigns**

	Campaign 1	Campaign 2	Both campaigns
Time period covered	Early fall '07	Late fall '07	
Length of campaign	14 days	10 days	
Number of ads displayed	32,272,816	9,664,332	41,937,148
Number of users shown ads	814,052	721,378	867,839
Treatment group viewing ads	63.7%	56.5%	67.9%
Mean ad views per viewer	39.6	13.4	48.3

Source: Lewis and Reiley (2013a).

The treatment group of 1.3 million Yahoo! users was exposed to two different advertising campaigns over the course of two months in fall 2007, separated by approximately one month. Table 7.1 gives summary statistics for the campaigns, which delivered 32 million and 10 million impressions, respectively. The two campaigns exposed ads to a total of 868,000 users in the 1.3-million-person treatment group. These individuals viewed an average of forty-eight ad impressions per person.

The experiment indicated an increase in sales of nearly 5 percent relative to the control group during the campaign, a point estimate that would translate to an extremely profitable campaign (with the retailer receiving nearly a 100 percent rate of return on the advertising spending). However, purchases had sufficiently high variance (due in part to 95 percent of consumers making zero purchases in a given week) to render the point estimate not statistically significantly different from zero at the 5 percent level. Controlling for available covariates (age, gender, state of residence) did not meaningfully reduce standard errors. This is a good example of how economically important effects of advertising can be statistically very difficult to detect, even with a million-person sample size. Just as we saw in section 7.2, we see here that the effects of advertising are so diffuse, explaining such a small fraction of the overall variance in sales, that the statistical power can be quite low. For this experiment, power calculations show that assuming the alternative hypothesis that the ad broke even is true, the probability of rejecting the null hypothesis of zero effect of advertising is only 21 percent.

The second important result of this initial study was a demonstration of the biases inherent in using cross-sectional econometric techniques when there is endogenous advertising exposure. This is important because these techniques are often employed by quantitative marketing experts in industry. Abraham (2008), for example, advocates comparing the purchases of exposed users to unexposed users, despite the fact that this exposure is endogenously determined by user characteristics and browsing behavior, which might easily be correlated with shopping behavior. To expose the biases in these methods, LR temporarily "discarded" their control group and compared the levels of purchases between exposed and (endogenously) unexposed parts of the treatment group. The estimated effects of advertising were three times as large as in the experiment, and with the opposite sign! This erroneous result would also have been deemed highly statistically significant. The consumers who browsed Yahoo! more intensely during this time period (and hence were more likely to see ads) tended to buy less, on average, at the retailer, regardless of whether they saw the ads or not (this makes sense, because as we will see most of the ad effect occurred offline). The control group's baseline purchases prior to the ad campaign showed the same pattern. Without an experiment an analyst would have had no way of realizing the extent of the endogeneity bias (in this case, four times as large as the true causal effect size) and may have come to a strikingly wrong conclusion.

Observing the consistent differences between exposed and unexposed groups over time motivated LR to employ a difference-in-differences estimator. Assuming that any unobserved heterogeneity was constant over time allowed LR to take advantage of both exogenous and endogenous sources of variation in advertising exposure, which turned out to reduce standard errors to the point where the effects were statistically significant at the 5 percent level. The point estimate was approximately the same as (though slightly higher than) the straight experimental estimate, providing a nice specification check. With this estimator, LR also demonstrated that the effects of the advertising were persistent for weeks after the end of the campaign, that the effects were significant for in-store as well as online sales (with 93 percent of the effect occurring offline), and that the effects were significant even for those consumers who merely viewed but never clicked the online ads (with an estimated 78 percent of the effect coming from nonclicking viewers). In a companion paper (Lewis and Reiley, forthcoming), the authors also showed that the effects were particularly strong for the older consumers in the sample—sufficiently strong to be statistically significant even with the simple (less efficient) experimental estimator.

In a follow-up study, Johnson, Lewis, and Reiley ([2013], henceforth JLR) improved on some of the weaknesses of the design of the original LR experiment. First, JLR ran "control ads" (advertising one of Yahoo!'s own services) to the control group, allowing them to record which control-group members would have been exposed to the ad campaign if they had been in the treatment group. This allowed them to exclude from their analysis those users (in both treatment and control groups) who were not exposed to the ads and therefore contributed noise but no signal to the statistics. Second, JLR convinced the advertiser to run equal-sized treatment and control groups, which improved statistical power relative to the LR article's 81:19 split. Third, JLR obtained more detailed data on purchases: two years of precampaign sales data on each individual helped to explain some of the variance in purchases, and disaggregated daily data during the campaign allowed them to exclude any purchases that took place before the first ad delivery to a given customer (which, therefore, could not have been caused by the ads, so including those purchases merely contributed noise to the estimates). The more precise estimates in this study corroborate the results of LR, showing point estimates of a profitable 5 percent increase in advertising, which are statistically significant at the 5 percent level, though the confidence intervals remain quite wide.

7.5 Activity Bias

In the preceding sections, we have presented this argument on an abstract level, arguing that the since the partial R^2 of advertising, even for a successful campaign, is so low (on the order of 0.00001 or less), the likelihood of omitted factors not accounting for this much variation is unlikely, especially since ads are targeted across time and people. In this section we show that our argument is not just theoretical. Here we identify a bias that we believe is present in most online ad serving; in past work, we gave it the name "activity bias" (Lewis, Rao, and Reiley 2011). Activity bias is a form of selection bias based on the following two features of online consumer behavior: (1) since one has to be browsing online to see ads, those browsing more actively on a given day are more likely to see your ad; and (2) active browsers tend to do more of *everything online*, including buying goods, clicking links, and signing up for services. Any of the selection mechanisms that lead to their exposure to the advertising are highly correlated with other online activities. Indeed, many of the selection mechanisms that lead to their exposure to the advertising, such as retargeting[20] and behavioral targeting, are highly correlated with other online activities. Hence, we see that ad exposure is highly and noncausally correlated with many online activities, making most panel and time-series methods subject to bias. In a nonexperimental study, the unexposed group, as compared to the group exposed to an ad, typically failed to see the ad for one or both of the following reasons: the unexposed users browsed less actively or the user did not qualify for the targeting of the campaign. When the former fails, we have activity bias. When the latter fails, we have classic selection bias.

In our 2011 paper, we explored three empirical examples demonstrating the importance of activity bias in different types of web browsing. The first application investigates the causal effects of display ads on users' search queries. In figure 7.1 we plot the time series of the number of searches by exposed users for a set of keywords deemed to be brand-relevant for a firm. The figure shows results for a time period that includes a one-day-display advertising campaign for a national brand on www.yahoo.com.

The campaign excluded a randomized experimental control group, though for the moment we ignore the control group and focus on the sort

20. For a discussion and empirical analysis of retargeting see Lambrecht and Tucker (2013).

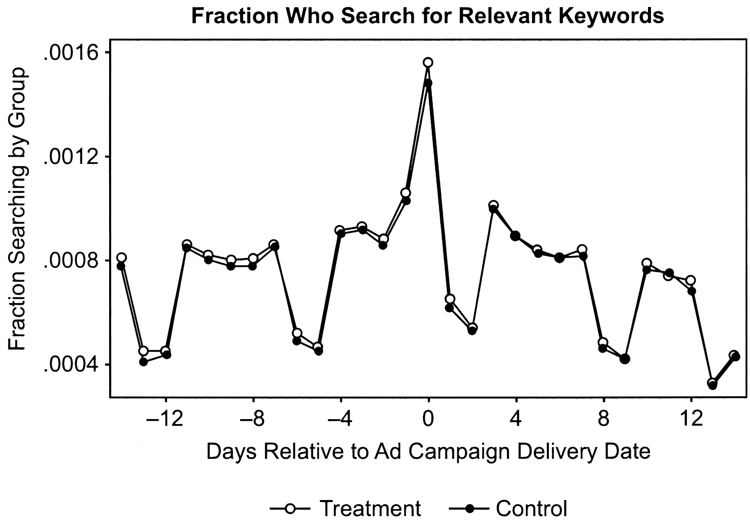

Fig. 7.1 Brand keyword search patterns over time
Source: Lewis, Rao, and Reiley (2011).

of observational data typically available to advertisers (the treatment group, those that saw the firm's advertisements). The x-axis displays days relative to the campaign date, which is labeled as Day 0. One can easily see that on the date of the ad, ad viewers were much more likely to conduct a brand-relevant search than on days prior or following. The advertising appears to *double* baseline search volume. Is this evidence of a wildly successful ad? Actually, no. Examining the control group, we see almost the same trend. Brand-relevant keyword searches also spike for those who saw a totally irrelevant ad. What is going on? The control group is, by design of the experiment, just as active online as the treatment group, searching for more of *everything*, not just the brand-relevant keywords of interest. The time series also shows that search volume is positively serially correlated over time and shows striking day-of-week effects—both could hinder observational methods. The true treatment-control difference is a statistically significant, but far more modest, 5.1 percent. Without an experiment, we would have no way of knowing the baseline "activity-related increase" that we infer from the control group. Indeed, we might have been tempted to conclude the ad was wildly successful.

Our second application involves correlation of activity not just across a publisher and search engine, but across very different domains. We ran a marketing study to evaluate the effectiveness of a video advertisement promoting the Yahoo! network of sites. We recruited subjects on Amazon Mechanical Turk, showed them the video, and gave them a Yahoo! cookie so we could track their future behavior. Using the cookie we could see if the

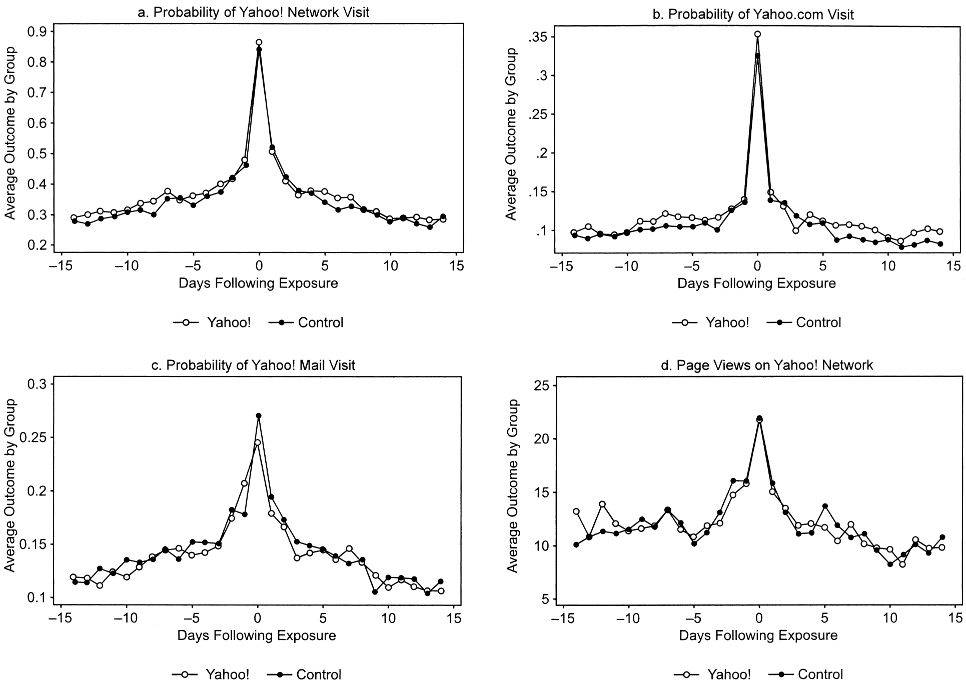

Fig. 7.2 The effect on various Yahoo! usage metric of exposure to treatment/control ads

Source: Lewis, Rao, and Reiley (2011).

Note: Panels A, B, and C: probability of at least one visit to the Yahoo! network, Yahoo.com, and Yahoo! mail, respectively. Panel D: total page views on the Yahoo! network.

ad really generated more Yahoo! activity. The control group saw a political ad totally unrelated to Yahoo! products and services. Again, we ignore the control group to begin. Figure 7.2 has the same format as figure 7.1 Day 0 on the x-axis labels the day an individual saw the video ad (with the actual calendar date depending on the day the subject participated in the study).

Examining the treatment group, we can see that on the day of and the days following ad exposure, subjects were much more likely to visit a Yahoo! site as compared to their baseline propensity, indicating a large apparent lift in engagement. However, data on the control group reveals the magnitude of activity bias—a very similar spike in activity on Yahoo! occurs on the day of placebo exposure as well. Both groups also show some evidence of positive serial correlation in browsing across days: being active today makes it more likely that you will be active tomorrow as compared to several days from now. People evidently do not engage in the same online activities (such as visiting Yahoo! and visiting Amazon Mechanical Turk) every day, but they engage in somewhat bursty activity that is contemporaneously correlated across sites. Online activity leads to ad exposure, which mechanically tends to occur on the same days as outcome measures we hope to affect with advertising. In the absence of a control group, we can easily make errors in causal inference due to activity bias. In this particular case, the true causal effect of the ad was estimated to be small and not statistically significant—given the cost of running a video ad, it was probably not worth showing, but the biased estimates would have led us to a wrong conclusion in this regard.

The third application again involves multiple websites. This time the outcome measure was filling out a new account sign-up form at an online brokerage advertised on Yahoo! Finance. Again, our results show that even those who were randomly selected to see irrelevant placebo ads were much more likely to sign up on the day they saw the (placebo) ad than on some other day. We refer the reader to our original paper for the details, stating here that the results are very similar to the ones we have just presented (the now familiar mountain-shaped graphs are again present). With activity bias it seems that one could erroneously "show" that nearly any browsing behavior is caused by nearly any other browsing behavior! We hope that our results will cause industry researchers to be more cautious in their conclusions. Activity bias is a real form of bias that limits the reliability of observational methods.

In the absence of an experiment, researchers may be able to use some other cross-validation technique in order to check the robustness of causal effects. For example, one could measure the effect of movie advertisements on searches for the seemingly irrelevant query "car rental." Similarly, one could check whether (placebo) ad views of a Toyota ad on the *New York Times* website on May 29 causes the same effect on Netflix subscriptions that day as did the actual Netflix ad on the *New York Times* website on May 30. Differences in differences using such pseudo-control groups will likely give better estimates of true causal effects than simple time-series or cross-sectional studies, though, of course, a randomized experiment is superior if it is available (Lewis, Rao, and Reiley 2011).[21]

Is activity bias a new phenomenon that is unique to the online domain? While it is not obvious that offline behavior is as bursty and as contemporaneously correlated as online behavior, before our study we did not think these patterns were obvious in online behavior either (and scanning industry white papers, one will see that many others still do not find it obvious!). We believe the importance of activity bias in the offline domain is an open question. It is not difficult to come up with examples in which offline advertising exposure could spuriously correlate with dependent variables of interest. Billboards undoubtedly "cause" car accidents. Ads near hospitals "cause" illness. Restaurant ads near malls probably "cause" food consumption in

21. In some cases, even such placebo tests may fail as the qualifications for seeing the ad may be intrinsically correlated with the desired outcome, as may be the case for remarketing and other forms of targeting, which account for search activity and browsing behavior.

general. Exposure to ads in the supermarket saver are likely correlated with consumption of unadvertised products, and so forth. The superior quality of data (and experiments) available in online advertising has laid bare the presence of activity bias in this domain. We believe the level of activity bias in other domains is an interesting, open question.

7.6 Measuring the Long-Run Returns to Advertising

Any study of advertising effectiveness invariably has to specify the window of time to be included in the study. While effects of advertising could in principle last a long time, in practice one must pick a cut-off date. From a business perspective, making decisions quickly is an asset worth trading decision accuracy for at the margin. But can patient scholars (or firms) hope to measure the long-run effects of advertising? Here we address the statistical challenges of this question. The answer, unfortunately, is rather negative. As one moves further and further from the campaign date, the cumulative magnitude of the sales impact tends to increase. (This is not guaranteed, as ads could simply shift purchases forward in time, so a short time window could measure a positive effect while a long time window gives a zero effect. But in practice, we have so far noticed point estimates of cumulative effects to be increasing in the time window we have studied.) However, the amount of noise in the estimate tends to increase faster than the increase in the signal (treatment effect) itself because in the additional data the control and treatment groups look increasingly similar, making long-run studies less statistically feasible than short-run ones. In the remainder of this section we formalize and calibrate this argument.

We again employ the treatment versus control t-statistic indexed by little t for time. For concreteness, let time be denominated in weeks. For notational simplicity, we will assume constant variance in the outcome over time, no covariance in outcomes over time,[22] constant variance across exposed and unexposed groups, and balanced group sizes. We will consider the long-term effects by examining a cumulative t-statistic (against the null of no effect) for T weeks rather than a separate statistic for each week. We write the cumulative t-statistic for T weeks as:

$$t_{\Delta\bar{y}t} = \sqrt{\frac{N}{2}}\left(\frac{\sum_{t=1}^{T}\Delta_{\bar{y}t}}{\sqrt{T}\,\hat{\sigma}}\right). \tag{8}$$

At first glance, this t-statistic appears to be a typical $O(\sqrt{T})$ asymptotic rate with the numerator being a sum over T ad effects and the denominator grow-

22. This assumption is clearly false: individual heterogeneity and habitual purchase behavior result in serial correlation in purchasing behavior. However, as we are considering the analysis over time, if we assume a panel structure with fixed effect or other residual-variance absorbing techniques to account for the source of this heterogeneity, this assumption should not be a first-order concern.

ing at a $\sqrt{T}$ rate. This is where economics comes to bear. Since $\Delta_{\bar{y}t}$ represents the impact of a given advertising campaign during and following the campaign (since $t = 1$ indexes the first week of the campaign), $\Delta_{\bar{y}t} \geq 0$. But the effect of the ad each week cannot be a constant—if it were, the effect of the campaign would be infinite. Thus, it is generally modeled to be decreasing over time.

With a decreasing ad effect, we should still be able to use all of the extra data we gather following the campaign to obtain more statistically significant effects, right? Wrong. Consider the condition necessary for an additional week to increase the t-statistic:

$$t_{\Delta\bar{y}T} < t_{\Delta\bar{y}T+1}$$

$$\frac{\sum_{t=1}^{T}\Delta_{\bar{y}t}}{\sqrt{T}} < \frac{\sum_{t=1}^{T+1}\Delta_{\bar{y}t}}{\sqrt{T+1}}.$$

Some additional algebra leads us to

$$1 + \frac{1}{T} < \left(1 + \frac{\Delta_{\bar{y}T+1}}{\sum_{t=1}^{T}\Delta_{\bar{y}t}}\right)^2,$$

which approximately implies

$$\text{(9)} \qquad \frac{1}{2}\cdot\frac{1}{T}\sum_{t=1}^{T}\Delta_{\bar{y}t} < \Delta_{\bar{y}T+1}.$$

This last expression says, "If the next week's expected effect is less than one-half the average effect over all previous weeks, then adding it in will only reduce precision." Thus, the marginal week can easily cloud the previous weeks, as its signal-to-noise ratio is not sufficiently large enough to warrant its inclusion.[23] If the expected impact of the campaign following exposure decays rapidly (although not necessarily all the way to zero), it is likely that including additional weeks beyond the campaign weeks will decrease the statistical precision.

Suppose that you were just content with the lower bound of the confidence interval increasing in expectation. A similar calculation, under similar assumptions, shows that the lower bound of a 95 percent confidence interval will increase if and only if

$$\text{(10)} \qquad 1.96(\sqrt{T+1} - \sqrt{T}) < \frac{\Delta_{\bar{y}T+1}}{\hat{\sigma}/\sqrt{N}}$$

where the right-hand expression is the marginal expected t-statistic of the $T + 1^{th}$ week.

23. Note that this expression is completely general for independent random draws under any marginal indexing or ordering. In the identically distributed case, though, the expected mean for the marginal draw is equal to all inframarginal draws, so the inequality always holds.

We can summarize these insights by returning to our formula for the t-statistic:

$$t_{\Delta \bar{y} T} = \sqrt{\frac{N}{2}} \left(\frac{\sum_{t=1}^{T} \Delta_{\bar{y}t}}{\sqrt{T}\sigma} \right).$$

Since the denominator is growing at $O(\sqrt{T})$, in order for the t-statistic to grow, the numerator must grow at a faster rate. In the limit we know this cannot be, as the total impact of the advertising would diverge faster than even the harmonic series.[24]

Now, ex ante it is hard to know when the trade-off turns against you. The effect may decay slower than the harmonic series initially and then move toward zero quite quickly. Of course, if we knew the pattern of decay, we would have answered the question the whole exercise is asking! So in the end, the practitioner must make a judgment call. While choosing longer time frames for advertising effectiveness analyses should capture more of the cumulative effect (assuming that it is generally positive), including additional weeks may just cloud the picture by adding more noise than ad impact. Measuring the effects of advertising inherently involves this sort of "judgment call"—an unsatisfying step in the estimation process for any empirical scientist. But the step is necessary since, as we have shown, estimating the long-run effect of advertising is a losing proposition—the noise eventually overwhelms the signal. The question is "when," and right now our judgment call is to use one to four weeks, but this is far from the final word.

7.7 Advances in Computational Advertising

In traditional media, targeting is typically a human-controlled process of determining the demographic groups most likely to consume the product. Readers may be familiar with Nielsen ratings for television, which break down to viewership by demographic categories. Campaigns often have "reach goals" for specific demographics a firm is interested in advertising to and marketing representatives use a portfolio of media outlets to meet these goals.

Online advertising opens up the possibilities for automated approaches to targeting because online ad delivery systems both gather information about specific users and make real-time, ad-serving decisions. "Computational advertising" is described by one of the founders of the field, Andrei Broder, as "a principled way to find the best match between a given user in a given context and a suitable advertisement" (Broder 2008). In traditional media, you have to specify who you want to advertise to. With computational adver-

24. We note that an asset with infinite (nominal) returns is not implausible per se (a consol does this), but we do find infinite effects of advertising implausible. The harmonic series is $\sum(1/t)$ whereas the requisite series for an increasing t-statistic would be $\approx \sum(1/\sqrt{t})$, which diverges much more quickly.

tising, you instead specify outcome metrics—an end-goal supported by the system—and the system's algorithms determine how to achieve that goal most efficiently. The end goal could be online sign-ups, clicks to a sales page, and so on. The end goals a system can support is limited by the bidding rules and data feedback supported by the advertising exchange. Some supported goals, such as conversions, might exhibit slow learning because the success rate is so low (1 in 300,000 would not be uncommon for account sign-ups, for instance).

While the details of these systems are well beyond the scope of this chapter, we will give the flavor of how they work. Which display ad to show can be modeled as a multiarmed bandit problem. The possible ads are the "arms" and a user-ad pair is a "pull of the arm." Papers in this literature adapt classic machine learning tools to the ad-serving context (see, for instance, Pandey, Chakrabarti, and Agarwal [2007]). A complimentary approach (which borrows from search advertising technology) is to view the advertisement as a document that must be retrieved and matched to the content page the ad is served on (which can be thought of as the query in search terminology; Rusmevichientong and Williamson 2006; Cary et al. [2007]).

We view (the current incarnation of) computational advertising primarily as automated targeting and local bid adjustment (to equalize CTR across campaigns, for instance). It helps to *locally* optimize ad spend by minimizing costs for given a campaign goal. By using an end goal, such as clicks or sign-ups, combined with a budget, these systems reduce the need to set targeting dimensions (reduces, not eliminates, because one might still set priors for the learning system, which might matter a lot in slow-to-learn tasks) and funnels spend to better performing inventory. Focusing on end goals also helps shift the conversation from "Who should get ads?" to "What do we want to get from our ad spending?" Practitioners should be cautious, however, that the system does not conflate "the audience most likely to convert" to the "audience that delivers the most *additional* conversions." To see the difference, imagine a customer that would buy anyway, but finds it convenient to click on an ad if he sees one. Paying for this conversion is a total waste of money. In our experience, some automated systems fail to draw this distinction and in doing so "order anticipate" by advertising to people likely to make a future purchase anyway. A natural solution is to integrate computational advertising with experimental platforms to provide randomization in order to measure incremental conversions. The technical infrastructure to make this possible would require advertisers to express a demand for reliable information.

Computational advertising is providing advances in advertising science. It can improve efficiency in the market by providing a better match of advertisements to consumers, thereby creating value, but current systems do not solve all the challenges we have laid out thus far, such as how much of a given action should be attributed to a given ad. For instance, suppose an online brokerage calculates that it nets $100 in profit from every account sign-up. Should it specify $100 as a maximum bid on an automated system and then "set it and forget it?" Presumably the brokerage is advertising heavily on TV and other media, including other online media that was not the "last click." Bidding $100 effectively says all this other spending gets zero credit—the firm would overadvertise using this rule. Of course, this is just the attribution problem reframed from the advertiser's perspective.

Thus, it is our opinion that many of the difficulties we have discussed about globally optimizing ad spending apply to *the current incarnation* of computational advertising as well. Perhaps the next revolution in advertising science will be in core algorithms to conduct automated experiments to measure incremental conversions and self-govern bids based on the experimental feedback. To our knowledge, there are no major technical barriers to this sort of pervasive experimentation and it has been applied fruitfully to infer causality in other online settings (Li et al. 2010). The challenge is that unlike ranking in search or recommendation of a news story, the response rate on a profitable ad is very low, on the order of 1/100–1/1000 for clicks on a display ad and an order of magnitude smaller for purchases, meaning feedback typically has low informational content. A second challenge is that the advertising exchange would have to facilitate the use of this technology by providing data on auctions the advertiser did not win (due to randomly entering a bid of 0 to experiment, for instance), which interacts with privacy concerns and platform incentives in interesting ways. Current practice does not provide this level of feedback, and we discuss workarounds firms currently use further on.

7.8 Moving Forward

Digital measurement has opened up many doors in measuring advertising effectiveness, but many challenges persist. In this section we look toward the future and discuss how we think many of the existing challenges will be overcome. Overall, we expect the advances to mainly come from better experimentation infrastructure to generate high-quality data at scale.

Experimental infrastructure has the potential to drastically reduce the cost of experimentation. The first generation of field experiments we ran at Yahoo! randomly selected a relatively small sample of users targeted by the campaign to see an unrelated advertisement. The problem was that an unrelated ad had to be entered into the booking system and run for the users that were randomized into the control group. The booking system was set up so that a firm could run multiple "creatives" (different versions of the ad) and the firm for whom we ran the experiment did not want to let another retailer get the traffic, because the competitor would benefit from the targeting dimensions set up by the retailer (including, for instance, past purchasing behavior).[25] The solution was to use charity ads for the control group. But this meant that either the advertiser had to pay for the control ads or Yahoo! had to donate them—both options came at a cost that increased linearly in the size of the control group, meaning that first generation experiments had relatively small control groups.

A small control group not only hurts power but also makes experimentation less useful as an evaluative tool. An experiment with 90 percent of subjects in the treatment group and 10 percent in the control has the same power as one with 10 percent in the treatment and 90 percent control. If control ads are free, then an advertiser could run nine of the latter for the cost of one of the former.[26] For control ads to be free, the ad server needs to be able to serve the "next ad in line" every time a user is randomized into the control group. Technologically, this requires a short-serving latency between the request to the ad server, the randomization, and the request for the replacement ad. The replacement ads are known as "ghost ads"—ads that naturally qualified to be served to a given user targeted by the campaign under study but not associated with the advertiser. Ghost ads make exploration and evaluation cheaper. Small treatment groups limit cost and allow advertisers to hone copy early in a campaign, while free control subjects help evaluate the campaign ex post.

Major online publishers are developing similar experimentation platforms. As experiments become cheaper and easier to run, advertisers will be able to form more precise beliefs on effectiveness than has heretofore been possible and further integrate experimentation into computational advertising platforms. These systems could incorporate an informative prior, which would help combat the power concerns we detailed earlier.

Another experimentation technology that improves power is the preexperiment matching of users. To see how this works, consider an experiment with subjects spread across treatment and control fifty-fifty. A standard experiment would simply flip a coin each time a user arrived at the website and show the ad corresponding to the outcome of the flip. Matching works as follows: Specify a set of attributes you care about such as recent sales levels and linear time trend. Form pairs of users by minimizing some objective function that defines the distance between two nodes in the graph of users. Then for each pair, flip a coin to determine experimental grouping. By

25. The treatment/control comparison would also provide the answer to a different question of advertising effectiveness.

26. Note that the statistical gains from such a change in experimental design are threefold. Further altering the design, assuming constant returns to scale from advertising (Lewis 2010; Johnson, Lewis, and Reiley 2012), by concentrating the 90 percent treatment group's ad impressions all within a smaller 10 percent treatment group expects an impact that is nine times as large, resulting in the equivalent ad effectiveness insights from running 81 of the 90 percent/10 percent experiments, producing confidence intervals of the ROI that are nine times more precise at no additional advertising cost.

construction, the specified metrics should be almost exactly equal between the two groups. For evaluating a noisy variable such as sales, guaranteeing the preperiod sales were the same can be useful. The treatment assignment is still totally exogenous, so all our normal intuition on how experiments identify causal effects goes through. Recent work has demonstrated that these techniques can double the power of experiments in many relevant settings (Deng, Kohavi, and Walker 2013).

These experimentation technologies create great potential for the next generation of computational advertising algorithms that we discussed in the last section. Automated experimentation would not be possible without the ability to deliver "non-ads" for free, record the interaction, and provide this feedback to nonwinning bidders. Of course, major publishers and exchanges will have to facilitate this capability, currently an advertiser bidding on an exchange only gets data on the impression (and what happens to the user) only when it wins the auction. Temporally (or geographic) based experiments offer something of a workaround, but can severely damage power (Nosko, and Tadelis 2014). As to whether this capability becomes standard practice for ad exchanges will presumably depend on advertiser demand, market power by major ad exchanges, and privacy legislation.

The future is also looking up for evaluating television advertising and associated "cross media" interaction effects (Joo et al. 2013). More people are viewing TV through devices like the Xbox and through services like Google TV, both of which link users to ads in systems similar to major web publishers. Furthermore, these users often have identifiers that can link television, sponsored search, and display ads for a single individual. Never before in the history of advertising has this been possible. The ability to measure cross-channel effects with the reliability of randomized experiments opens the door to many new questions for academics and many new strategies for advertisers. As more forms of advertising become measurable on an individual level, our ability to provide reliable estimates of advertising effectiveness will expand as well. The advances so far have already set a new state of the art in measurement, and we expect the trend to continue.

7.9 Concluding Remarks

The science of measuring advertising effectiveness has evolved considerably due to new digital data sources and experimentation platforms. We view experimentation on the individual level with the ad delivery linked to purchasing behavior as a true game changer offered by digital media as compared to traditional counterparts. Whether in search or display, new advertisers can gather feedback that is immune from the biases that plague observational methods. Another important advance is computational advertising. Computational advertising helps solve the targeting problem and usefully shifts the conversation from "who to hit" to "what do I get." Yet neither of these advances has yet to solve all the measurement problems in advertising science. Experiments are noisy and computational advertising still relies on humans to enter the key parameters, such as valuations of clicks or conversions, that govern spend. The future holds promise, but depends on economic incentives that at this point are hard to predict.

Moving forward, experimentation and data collection technology is evolving alongside new forms of ad serving and computational advertising systems. Questions such as the cross-derivative of certain media on the effectiveness of other media will be in play in the coming years. Measuring the effectiveness of media, such as television, that were previously not technologically feasible because randomizing delivery was not possible at scale, will also greatly expand knowledge on advertising effectiveness. This will in turn allow firms to more accurately guide their advertising expenditure. Our view, however, is that challenges such as measuring the long-run effects of advertising and the impact of brand advertising appear to be out of reach for at least the next five to ten years, if not longer. We await new developments in advertising science at the digital frontier to facilitate the answers to these and new questions.

References

Abraham, M. 2008. "The Off-Line Impact of Online Ads." *Harvard Business Review* 86 (4): 28.

Bagwell, K. 2007. "The Economic Analysis of Advertising." In *Handbook of Industrial Organization*, vol. 3, edited by M. Armstrong and R. Porter. Amsterdam: North Holland.

Becker, G. S., and K. M. Murphy. 1993. "A Simple Theory of Advertising as a Good or Bad." *Quarterly Journal of Economics* 108 (4): 941–64.

Blake, T., C. Nosko, and S. Tadelis. 2014. "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment." NBER Working Paper no. 20171, Cambridge, MA.

Broder, A. 2008. "Computational Advertising and Recommender Systems." In *Proceedings of the 2008 ACM Conference on Recommender Systems*, 1–2. Association for Computing Machinery.

Bureau of Labor Statistics. 2010. *American Time Use Survey*. http://www.bls.gov/tus/charts/leisure.htm.

Card, D. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, vol. 3, edited by O. C. Ashenfelter and D. Card, 1801–63. Amsterdam: North Holland.

Cary, M., A. Das, B. Edelman, I. Giotis, K. Heimerl, A. Karlin, C. Mathieu, and M. Schwarz. 2007. "Greedy Bidding Strategies for Keyword Auctions." In *Proceedings of the 8th ACM conference on Electronic Commerce*, 262–71. Association for Computing Machinery.

Chan, D., R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. 2010. "Evaluating Online Ad Campaigns in a Pipeline: Causal Models at Scale." In *Proceedings of*

the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 7–16. Association for Computing Machinery.
Demsetz, H. 1982. "Barriers to Entry." *American Economic Review* 72 (1): 47–57.
Deng, A., Y. Xu, R. Kohavi, and T. Walker. 2013. "Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data." In *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*, 123–32. Association for Computer Machinery.
Dickie, M., A. Fisher, and S. Gerking. 1987. "Market Transactions and Hypothetical Demand Data: A Comparative Study." *Journal of the American Statistical Association* 82 (397): 69–75.
Gelman, A., and J. Carlin. 2013. "Beyond Power Calculations to a Broader Design Analysis, Prospective or Retrospective, Using External Information." Working Paper, Columbia University.
Goldfarb, A., and C. Tucker. 2011a. "Online Display Advertising: Targeting and Obtrusiveness." *Marketing Science* 30 (3): 389–404.
———. 2011b. "Search Engine Advertising: Channel Substitution When Pricing Ads to Context." *Management Science* 57 (3): 458–70.
Gonen, R., and E. Pavlov. 2007. "An Incentive-Compatible Multi-Armed Bandit Mechanism." In *Proceedings of the 26th Annual ACM Symposium on Principles of Distributed Computing*, 362–3. Association for Computing Machines.
Hu, Y., L. M. Lodish, and A. M. Krieger. 2007. "An Analysis of Real World TV Advertising Tests: A 15-Year Update." *Journal of Advertising Research* 47 (3): 341.
Johnson, G., R. A. Lewis, and D. H. Reiley. 2012. "Location, Location, Location: Geo-Targeting Increases Effectiveness of Online Display Advertising." Unpublished manuscript.
———. 2013. "Add More Ads? Experimentally Measuring Incremental Purchases Due to Increased Frequency of Online Display Advertising." Working Paper, Rochester University.
Joo, M., K. C. Wilbur, B. Cowgill, and Y. Zhu. 2013. "Television Advertising and Online Search." *Management Science* 60 (1): 56–73.
Kessides, I. N. 1986. "Advertising, Sunk Costs, and Barriers to Entry." *Review of Economics and Statistics* 68 (1): 84–95.
Kumar, D., and T. Yildiz. 2011. "Measuring Online Ad Effectiveness." In *12th ACM Conference on Electronic Commerce*. Association for Computing Machines.
Lambrecht, A., and C. Tucker. 2013. "When Does Retargeting Work? Information Specificity in Online Advertising." *Journal of Marketing Research* 50 (5): 561–76.
Lewis, R. 2010. "Where's the Wear-Out?" Online Display Ads and the Impact of Frequency." PhD diss., Massachusetts Institute of Technology.
Lewis, R. A., and D. T. Nguyen. 2013. "A Samsung Ad and the iPad: Display Advertising's Spillovers to Online Search." Unpublished manuscript.
Lewis, R. A., and J. M. Rao. 2013. "On the Near-Impossibility of Measuring the Returns to Advertising." Unpublished manuscript.
Lewis, R., J. Rao, and D. Reiley. 2011. "Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising." In *Proceedings of the 20th International Conference on World Wide Web*, 157–66. Association for Computing Machines.
Lewis, R. A., and D. H. Reiley. Forthcoming. "Advertising Effectively Influences Older Users: A Yahoo! Experiment Measuring Retail Sales." *Review of Industrial Organization*.
———. 2013a. "Online Advertising and Offline Sales: Measuring the Effects of Retail Advertising via a Controlled Experiment on Yahoo!" Unpublished manuscript.

———. 2013b. "Super Bowl Advertising Causes Down-to-the-Minute Online Search Behavior." In *Proceedings of the 14th ACM Conference on Electronic Commerce*. Association for Computing Machines.

Lewis, R., D. Reiley, and T. Schreiner. 2012. "Ad Attributes and Attribution: Large-Scale Field Experiments Measure Online Customer Acquisition." Unpublished manuscript.

Li, L., W. Chu, J. Langford, and R. E. Schapire. 2010. "A Contextual-Bandit Approach to Personalized News Article Recommendation." In *Proceedings of the 19th International Conference on World Wide Web*, 661–70. Association for Computing Machines.

Lodish, L., M. Abraham, S. Kalmenson, J. Livelsberger, B. Lubetkin, B. Richardson, and M. Stevens. 1995. "How TV Advertising Works: A Meta-Analysis of 389 Real World Split Cable TV Advertising Experiments." *Journal of Marketing Research* 32 (2): 125–39.

Lovell, M. 2008. "A Simple Proof of the FWL Theorem." *Journal of Economic Education* 39 (1): 88–91.

Mincer, J. 1962. "On-the-Job Training: Costs, Returns, and Some Implications." *Journal of Political Economy* 70 (5): 50–79.

Murphy, J., P. Allen, T. Stevens, and D. Weatherhead. 2005. "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation." *Environmental and Resource Economics* 30 (3): 313–25.

Pandey, S., D. Chakrabarti, and D. Agarwal. 2007. "Multi-Armed Bandit Problems with Dependent Arms." In *Proceedings of the 24th International Conference on Machine Learning*, 721–8. Association for Computing Machines.

Pandey, S., and C. Olston. 2006. "Handling Advertisements of Unknown Quality in Search Advertising." In *Advances in Neural Information Processing Systems*, edited by M. Jordan, Y. LeCun, and S. Solla, 1065–72. Cambridge, MA: MIT Press.

Papadimitriou, P., H. Garcia-Molina, P. Krishnamurthy, R. A. Lewis, and D. H. Reiley. 2011. "Display Advertising Impact: Search Lift and Social Influence." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1019–27. Association for Computing Machines.

Rusmevichientong, P., and D. Williamson. 2006. "An Adaptive Algorithm for Selecting Profitable Keywords for Search-Based Advertising Services." In *Proceedings of the 7th ACM Conference on Electronic Commerce*, 260–69. Association for Computing Machines.

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

Tucker, C. E. 2012. "The Economics of Advertising and Privacy." *International Journal of Industrial Organization* 30 (3): 326–29.