Discussion of Gordon et al. (2017):
“A Comparison of Approaches to Advertising Measurement:
Evidence from Big Field Experiments at Facebook

Daniel Zantedeschi*
The Ohio State University

April 2, 2018

1 Summary

First, I thank the QME editorial board for the opportunity to comment on Gordon et al. (2017) in writing. Gordon et al. (2017) are in a privileged condition, a “dream” for scholars interested in digital marketing: they can analyze multiple online advertising campaigns run on the popular social media and networking services company.

Firms bid for ads placement on the individuals’ Facebook pages using an auction mechanism. If successful in the auction, the firm has the option to perform standard A/B testing: a coin flip determines whether the ad is dispatched to the individual’s Facebook’s page, or if another firm’s ad is instead sent to the individual as a “control”. Importantly, the authors have access to individuals’ funnel information: whether the ad was visualized by the individual, intermediate (i.e., clicks and website landings), and

*Assistant Professor, Department of Marketing and Logistics, Fisher College of Business and Translational Data Analytics, The Ohio State University. 2100 Neil Avenue, 500 Fisher Hall, 43210 Columbus, Ohio, USA. Please direct correspondence to mailto:zantedeschi.1@osu.edu.
ultimate conversions, (i.e., purchases).

In this work “the campaign lift” is measured “on the treated,” namely, conditioning on the ad exposure detected by the web browser. This approach departs from the traditional marketing effectiveness literature (see Lodish et al., 1995), where estimates capture the “intention to treat.” 2

The focus of the paper is to provide estimates of the “lift” under two scenarios: (1) the experimental ideal taking advantage of the randomized assignment, and (2) an observational setup ignoring the presence of the randomized variation and attempting to address endogeneity concerns using historical information comprising detailed summaries of previous levels of social media activities and well-established statistical methodologies.

The authors find that observational methods resoundingly fail to recover their experimental counterpart. This result is robust to multiple campaigns run by different companies and has important repercussions for understanding the value of modern online ecosystems in which exposures are linked to intermediate and ultimate outcomes.

The substantive implication for advertisers running campaigns on social media is that randomized variation performed at the ad-placement level is the best safeguard against biased estimates of the campaigns lift. Furthermore, digital marketers should carefully consider the value of historical data (Rossi et al., 1996) when the aim is predicting individual-level behavior throughout the digital purchasing funnel3: the reason is that, consistent with what documented initially in Lewis et al. (2011), unobserved – and possibly unobservable – activities appear to drive the endogenous clicking/viewing behavior on Facebook.

Overall, this is a translational and impactful paper and it is deservedly cited as an example of “responsible data science” (van der Aalst et al., 2017). I am sure it will set the tone for further developments, both methodological and substantial, in digital marketing and will draw attention from both the academic circles and industry practitioners.

During the QME conference, I discussed more in-depth some of the technical nuances arising from
the statistical exercise. These are reported in the next Section together with comments for potential extensions or clarification.

2 Discussion on the Statistical Approaches

2.1 Analysis of the Experimental Data

2.1.1 ITT and ATT

Let $Y$ be a binary variable indicating the outcome under consideration (i.e. landing on the firm’s web page or actual purchase), $X$, also binary, identifies the treatment assignment, namely a banner ad being sent to an individual on Facebook, and let $W$ be a “compliance” binary variable indicating whether the ad was actually visualized as detected by the web browser. The information structure is represented in Figure (1).

![Figure 1](image)

Figure 1: $X$: Assignment indicator, 1 for treatment and 0 for control. $W$: Compliance indicator, 1 for compliant and 0 for defiant. $n_{1S}$ indicates the number of conversions in stratum $S$, where we observe three natural strata $S = \{0, 11, 01\}$. Similarly, $n_{0S}$ represents the number of non-conversions in the same stratum. As an important detail, the compliance status is not observed in the control group.

The average treatment effect, at “the intention to treat”, is given by:

$$ITT = E(Y(1) - Y(0))$$ (1)
where \( Y(X) \) indicates the outcome of the corresponding assignment, \( X = \{0, 1\} \). The attentive reader may note the ITT is directly identifiable from the data, as conversions counts are kept track of in each branch of the tree in Figure 1. Indeed, an unbiased estimate for the ITT is given by:

\[
\hat{ITT} = \frac{n_{111} + n_{101}}{n_{111} + n_{101} + n_{011} + n_{001}} - \frac{n_{10}}{n_{10} + n_{00}}
\]  

(2)

The “intention” is reflected in non-differentiating between the compliance and non-compliance strata when considering conversions. Essentially, whether the individual visualizes the ad or not is irrelevant from the firm’s perspective.

Gordon et al. (2017) depart from the ITT framework and leveraging on the information about the exposure \( W \), propose to consider the average effect on the treated, \( ATT \):

\[
ATT = E(Y(1) - Y(0)|W = 1).
\]  

(3)

A first suggestion for the authors is clarify the practical and managerial meaning of the \( ATT \). In the social sciences the \( ATT \) reflects the value of policies offered to individuals who opt into the “treatment” (e.g. a vocational training program) given participation incentives (e.g. an educational voucher). Hence, the \( ATT \) can be interpreted as the value of the policy amongst those who actively choose to participate into the program (Angrist, 2010). One may wonder whether there is an analogy in the online advertising ecosystem. From my perspective, it is somehow difficult to argue that individuals actively choose to visualize the ad sent to their Facebook page. For instance, the “visualization” of banner ads is determined by how web browsers optimize loading times affected by extraneous factors such as the size of the screen and the speed of the Internet connection.\(^4\) Hence, it is only partially under the control of the individual.

Moreover, the choice whether to opt for ITT or \( ATT \) estimates depends also on the objectives of the firm. For instance, if the aim is to compute the campaign ROI, and the contractual agreement between
the firm-agency and the publisher is based on the “attempt to reach,” then the ITT suffices. Finally, the ITT may be preferable when comparing the effectiveness of the campaigns longitudinally: there are peak times, (e.g. Black Friday to Cyber Monday) when advertisers flock to Facebook ad exchanges leading to increased crowdedness on individuals’ Facebook pages ultimately rendering visualization rates volatile across time.

2.1.2 The Exclusion Restriction

Passing to the econometrics, the first term in the definition of the ATT in (3), $E(Y(1)|W = 1)$, is directly identifiable by sampling variation: $\frac{n_{111}}{n_{111} + n_{011}} \to E(Y(1)|W = 1)$.

The second term in (3), $E(Y(0)|W = 1)$ is instead not directly linkable to the data and is a counterfactual: the expected outcome of those who were not sent the ad, had they been compliant. To solve this dilemma, Gordon et al. (2017) appeal to the exclusion restriction assumption (ER) stating that “the ad is effective only if there is an exposure”. Indeed, by the law of total probability:

\[
\begin{align*}
\text{ITT} &= E(Y(1) - Y(0)) = \\
&= E(Y(1) - Y(0)|W = 1) \Pr(W = 1) + E(Y(1) - Y(0)|W = 0) \Pr(W = 0)
\end{align*}
\]

Thus, under ER: $E(Y(1) - Y(0)|W = 0) = 0$ implies no difference in effectiveness under the treatment and control group whenever there is no exposure, $W = 0$. This simplification yields a “free lunch”:

\[
\text{ATT} = \frac{\text{ITT}}{\Pr(W = 1)}
\]

The formula (5) states that ATT is estimable from the observed data and is equivalent to the ITT up to a rescaling induced by the probability of exposure.\(^5\) Hence, an unbiased estimate at the ATT can be obtained by taking the ratio between $\hat{\text{ITT}}$ and an estimate of the probability of compliance, $\hat{\Pr}(W) = \ldots$
Importantly, the simplicity of the estimation of the \( ATT \) in (5), depends on the absence of uncertainty in the exposure counts. However, this is possibly not the case as the ad-delivery system implemented by Facebook may be prone to under/over counting the exposures.\(^6\) In light of these idiosyncratic uncertainties, as a potential extension, one could consider the approach developed in Frangakis and Rubin (2002), which does not strictly rely on ER. In short, Frangakis and Rubin (2002) consider the membership to the three strata \( S = \{0, 11, 01\} \) as probabilistic. The practical implication of this framework would be to increase the variance around the \( ATT \) estimates due to the effect of accounting for the uncertainty driving the stratum assignment. Thus, from my perspective, weakening the ER assumption could operate in the direction of “closing the gap” between the observational and experimental estimates due to the increased uncertainty in the latter.

2.2 The Observational Estimates

The authors proceed by assuming that there is no randomization at the ad-placement level and just consider the subset of the individuals to whom the ads were sent. This setup is represented in Figure 2. At

![Figure 2: Reduced-form tree ignoring the control group. The notation is the same as in figure 1 with the addition of historical information about the compliers and defiers denoted by \( Z_1 \) and \( Z_0 \) respectively.](image)

a quick glance, it would seem reasonable to ignore randomization and model the endogenous selection of individuals into the exposed category taking advantage of the large set of “pre-treatment” variables, \( Z \).
Without loss of generality, a two-level semi-parametric model can represent the compliance/conversion problem:

\[
\begin{align*}
\Pr(W = 1) &= f_W(Z, \varepsilon_W) \\
Y &= f_Y(Z, W, \varepsilon_Y)
\end{align*}
\]  

(6a)  

(6b)

where the unobserved variables are denoted as \(\varepsilon_W\) and \(\varepsilon_Y\). Equation (6a) is commonly referred as the **propensity score model**, while equation (6b) as the **outcome/response model**.

Gordon et al. (2017) attempt to estimate the \(\beta = \frac{\partial f_Y(Z, W)}{\partial W}\) (and the corresponding “lift”) from the system of equation in (6), as the natural estimator compatible with ATT in (5). In order to do so, a series of modeling assumptions and machine learning procedures are employed. These are summarized as follows:

- \(\varepsilon_W\) is assumed to be generalized extreme value distributed, leading to a logistic regression for (6a);
- (6a) is specified as a linear probability model;
- \(f_W\) and \(f_Y\) are both linearly and additive function of the covariates within the parentheses;
- Given the large dimensionality of the historical covariates, \(Z\), \(f_W(Z)\) and \(f_Y(Z, W)\) are “selected” using the Lasso regularizer, Tibshirani (1994). This, at least asymptotically, retains the “best” subset of the covariates based on predictive criteria (i.e. generalized cross validation).

The endogenous compliance problem is subsequently dealt with an array of well-established approaches belonging, loosely speaking, to two classes: (1) “adjustment” and (2) “weighting” methods. Importantly, these differ in the treatment of the information arising from the propensity score model in (6a).

In the “adjustment” method, the estimated propensity score \(\widehat{\Pr}(W = 1)\) is included as an additional covariate to account for the endogenous selection in the response equation, (6b). The intuition behind
is to “control/adjust”, as a baseline correction, for the endogenous compliance using as a “proxy” the estimated propensity score.

In the “weighting” method, the propensity score $\overline{PR(W = 1)}$ is instead used to either “discount/inflate” the observations in the likelihood of (6b) (Inverse probability weighting, IPW), or to find “twins” between the observations in terms of the score but experiencing different compliance behavior (Propensity score matching, PSM). The intuition behind the weighting method\(^7\) is instead to make the sample appear as if it was drawn from a random experiment based on $\overline{PR(W = 1)}$ \(^8\).

As mentioned earlier, despite the authors’ carefulness, these attempts fail to recover the estimates under the experimental conditions in the very large majority of the advertising campaigns. Interestingly, there appears to be no systematic pattern: the campaign lift is often overestimated, but other times underestimated. The question, then, arises: what went wrong?

2.2.1 Variable Selection Tools and Unobservables

While the Lasso criterion obtains (asymptotically) the best subset amongst the available covariates, there is still the possibility of unmeasured confounding: an unobservable variable affecting both the compliance and the response equation in (6).

The authors discuss the potential role of unobservables and perform an interesting sensitivity analysis in the spirit of Rosenbaum (1987) and directly linked to the approach developed in Ichino et al. (2008). The key finding is that to reconcile the differences between observational and experimental estimates a large amount of imputed noise is needed. This unobserved variation is akin to the so-called “activity bias” (Lewis et al., 2011): more active users tend to visualize ads more often, and more active users may be more (or less) likely to convert regardless.\(^9\)

A suggestion related to the statistical analyses would be to assess the practical and statistical value of the variable selection procedure. Specifically, I feel a discussion about which (classes of) variables are
retained could provide value to the paper. Moreover, while it seems reasonable to identify parsimonious models of the exposure and the outcomes, the Lasso could do *more harm than good*. For instance, the regularization procedure could remove variables that correlate with neither the exposure nor the outcome but that collectively could be associated with the unobserved variation determining the endogenous compliance\(^{10}\). Hence, a possible extension would be to benchmark against a *saturated* response model in (6b) including as many predictors and transformations of them (i.e. interactions) as possible Rossi (2014).

**Acknowledgments**

Comments from Greg Allenby, Anocha Aribarg, Thomas Otter, Edlira Shehu and Evan Weingarten on an early version of this discussion are duly appreciated. A special thank you to Brett Gordon and Wes Hartmann for the support and the patience throughout the process.
References


**Notes**

1. Namely, the second highest bidder in the auction. However, please note the second highest bidder need not be a direct competitor of the firm.

2. A notable exception in the digital marketing literature is provided by Johnson et al. (2017) where it is empirically demonstrated that “on the treated” estimates are more precise than their “intention to treat” counterpart.

3. For instance, in the design of segmentation, *behavioral* and *lookalike* targeting strategies.

4. The interested reader may refer to [https://varvy.com/pagespeed/display.html](https://varvy.com/pagespeed/display.html).

5. As also noted in the paper, formula (5) represents the ATT as a “Wald” estimator where $X$ is a perfect instrument for $W$.


7. IPW methods are commonly employed in primary data analysis in the presence of imbalance between the response classes.

8. Noting that the two categories are not exclusive, the authors also consider hybrid approaches that both “adjust” and “weigh” simultaneously (double-robust adjustments).


10. Please also refer to Armistead (2014) for a discussion about the role of “third variables” in controlling for unobservables and Montes-Rojas and Galvao (2014) who explored the interplay between variable selection and endogeneity bias.