Journal of the Royal Statistical Society Series A: Statistics in Society, 2026, 00, 1–21
https://doi.org/10.1093/jrsssa/qnag066
Published: 17 June 2026
Original article

The effect of political advertising after *Citizens United*: adjusting for unmeasured confounding in marginal structural models

Matthew Blackwell [1], **Soichiro Yamauchi** [2]

[1]Independent Scholar, Cambridge, MA 02138, USA
[2]Department of Political Science, University of California, San Diego, La Jolla, CA 92093, USA
Address for correspondence: Soichiro Yamauchi, Department of Political Science, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA. Email: soichiro@ucsd.edu

Abstract

Corporations, unions, and other interest groups have become key sponsors of television advertising in US elections after the Supreme Court's decision in *Citizens United v. FEC* that eliminated restrictions on such spending. This paper estimates the partisan effects of ads sponsored by these groups to obtain a more complete picture of voter behaviour and electoral politics. Advertising strategies vary over the course of the campaign, making marginal structural models a natural tool for this setting. Unfortunately, this approach requires an assumption of no unobserved confounders between the treatment and outcome, which may not be plausible with observational electoral data. We propose a novel weighting estimator with propensity-score fixed effects to adjust for time-constant unmeasured confounding in marginal structural models of fixed-length treatment histories. This estimator is consistent and asymptotically normal when the number of units and time periods grow at a similar rate. Unlike traditional fixed effect models, this approach works even when the outcome is only measured at a single point in time as in our setting. Against conventional wisdom, we find interest group ads are only effective when run by Democratic groups, and these effects are most prominent after Donald Trump became a presidential candidate in 2015.

1 Introduction

Television advertisements have been a cornerstone of US politics since they first aired in the 1950s (Benoit, 2013). In the 2019–2020 election cycle, there were 2.35 million ad airings in the presidential race, 2.33 million in Senate races, and 1.36 million in US House races, each number setting a record (Ridout et al., 2021). These ads are one of the main ways in which candidates, political parties, and outside interest groups attempt to influence voter behaviour, electoral outcomes, and, ultimately, public policy. A large literature in political science and related fields has attempted to estimate the effect of these ads on various outcomes and for various political offices (Blackwell, 2013; Goldstein & Ridout, 2004; Hill et al., 2013; Huber & Arceneaux, 2007; Jacobson, 1975; Ridout & Franz, 2011; Sides et al., 2022). The findings from these studies vary but generally point to ad airings having persuasive effects on voter behaviour with larger effects at lower levels of political office.

Received: July 11, 2024. **Revised:** February 10, 2026. **Accepted:** March 12, 2026

In this paper, we focus on the effects of advertisements sponsored by independent interest groups. The US Supreme Court decision in *Citizens United v. FEC* (2010) removed campaign finance restrictions on independent expenditures by outside interest groups, including corporations and labor unions. Since that time, the share of television ad airings sponsored by outside groups has grown from roughly 10% pre-*Citizens United* to over 25% in the period after the decision (Ridout et al., 2021). This growth has worried citizens and political observers since these groups were no longer required to disclose their donors, a phenomenon known as 'dark money', potentially providing a large benefit to corporations, wealthy individuals, and, ultimately, the Republican party. Indeed, early research into the effects of the ruling found that *Citizens United* increased Republican vote share in state legislative races (Klumpp et al., 2016), though these studies tended to focus on aggregate effects without focusing on advertising directly. Our goal is to measure how independent group television ads have affected US Senate and Gubernatorial races in the post-*Citizens United* era.

A major challenge to assessing the effects of political advertising is the dynamic nature of its deployment. Candidates and groups change the amount and content of advertising in response to how other groups advertise, which then affects the decisions of opponents. The feedback cycle of political advertising implies the potential for time-varying confounding that can bias our estimates of the effectiveness of advertising. Unfortunately, most studies of advertising ignore these issues and simply rely on aggregate campaign-level measures of advertising. One exception, Blackwell (2013), applied the combination of marginal structural models (MSM) and inverse probability of treatment weighting (IPTW) (Robins et al., 2000) to estimate the time-varying effects of negative advertising, showing considerable differences with estimates that ignore or poorly handle time-varying confounding. More generally, the use of marginal structural models for time-varying treatments in the social sciences has grown over the last few decades (Baćak & Karim, 2019; Bacak & Kennedy, 2015; Creamer & Simmons, 2019; Kurer, 2020; Ladam et al., 2018; Obikane et al., 2018; Sampson et al., 2006; Sharkey & Elwert, 2011; Wodtke et al., 2011).

One limitation of the IPTW approach to marginal structural models is that it usually relies on an assumption of sequential ignorability, which states that there are no unmeasured confounders between the treatment at time t and the outcome conditional on the treatment and covariate history up to that point. In social science studies, this assumption could be suspect when units select into treatment based on data not available to the researcher. In our setting, we might worry that groups will be more likely to advertise in certain media markets where voters are known by the campaigns to be favourable to the supported candidate. To overcome these issues, this article extends the IPTW approach to estimating the effects of time-varying treatments to allow for time-constant unmeasured confounding. To do so, we propose a straightforward modification to IPTW: to include unit-specific fixed effects in the propensity score model used to construct the inverse-probability weights. While this approach will lead to an incidental parameters problem for the propensity score model (Neyman & Scott, 1948), we show that if this model is correctly specified and the number of time periods grows at the same rate as the number of units, the IPTW with fixed effects estimator (IPTW-FE) will lead to a consistent and asymptotically normal estimator for the parameters of the marginal structural model. This is true even when we only have a single measurement of the outcome after the final instance of treatment, as is the case in our setting. This approach relies on a within-unit version of sequential ignorability, which allows the type of feedback between the treatment and outcome usually ruled out by linear outcome fixed effects estimators (Imai & Kim, 2019; Sobel, 2012). The essential logic of the IPTW-FE is quite simple. If the propensity score model is stable over time and we have a number of time periods, we can allow for each unit to have a unique offset to the propensity score model that should incorporate any time-constant variables, measured or unmeasured.

To prove our main results, we rely on a robust literature on nonlinear panel models that has established the asymptotic distribution of our propensity score estimator when the number of time periods grows at a similar rate to the number of units (Arellano & Hahn, 2007; Fernández-Val, 2009; Fernández-Val & Weidner, 2016, 2018; Hahn & Kuersteiner, 2011; Hahn & Newey, 2004). Many of these approaches have developed bias correction techniques since these estimators are often asymptotically biased. Our approach avoids this issue with these estimators for two reasons. First, we follow the MSM literature and focus on estimating the parameters of the MSM at the slower $\sqrt{N}$ rate rather than

the $\sqrt{NT}$ rate so that the asymptotic bias described in this literature converges to 0. Second, we focus on the effect of a finite number of lags of treatment, which limits how much the bias from noisy fixed effect estimation can affect the estimates of the MSM parameters.

Applying these methods to data on US Senate and Gubernatorial elections from 2010 to 2020, we find that each additional week of ads from independent groups supporting Democratic candidates increases the Democratic share of the two-party vote, increases Democratic turnout, and decreases Republican turnout. We find no such effects for ads from independent groups supporting Republican candidates, in spite of the conventional wisdom that spending from interest groups would generally favour Republicans. We additionally find that the effectiveness of pro-Democratic independent group ads is driven mostly by the post-2016 era after Donald Trump became a presidential candidate. Finally, we show that our method has null effects on a number of placebo tests, increasing our confidence that these results are not driven by unmeasured confounding.

Our methodological approach is also related to recent work on causal inference in fixed effects settings. Arkhangelsky and Imbens (2024) is most closely related to our approach here. They investigate how to use inverse probability weighting with fixed effects when a set of sufficient statistics for the treatment process is available, though in a fixed-T setting with no dynamic feedback between the treatment and the outcome and no time-varying covariates. Other work has explained how this dynamic feedback stymies estimation of both contemporaneous effects and the effects of treatment histories with fixed effects assumptions (Imai & Kim, 2019; Sobel, 2012). In contrast, our approach allows for feedback between the treatment and the outcome, so long as sequential ignorability holds conditional on the unit-specific effect. Finally, a large literature has grown recently to explain how and when difference-in-differences methods may be used to estimate the effects of time-varying treatments on outcomes when a panel of treatments and outcomes are observed together (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021; Sun & Abraham, 2021). In our application (and many others in the MSM literature), we only have a single endpoint measure of the outcome, so there are no 'pre-treatment' or baseline outcomes to leverage for removing unmeasured confounding.

The paper proceeds as follows. Section 2 introduces the data and notation for our setting. In Section 3, we review marginal structural models and inverse probability of treatment weighting as they are currently deployed in applied research. We then introduce our fixed-effect approach in Section 4, describing both the assumptions that justify its use and its large-sample properties under these assumptions. In Section 5, we present simulation evidence of the finite-sample performance of this estimator, which shows that it works well, especially when the amount of unmeasured heterogeneity is limited. Finally, we present our results in Section 6 and conclude with some ideas for future research in Section 7.

2 Data and notation

Our data consists of Senate and Gubernatorial general election campaigns in the United States from 2010 until 2020. These are state-wide races, but we analyse the data at the media market level, the lowest level at which we can obtain advertising data. Media markets consist of clusters of counties where a single group of broadcast television channels can reach. Our advertising data comes from the Wesleyan Media Project and contains political ads on all broadcast television stations in all media markets in the United States (Fowler et al., 2019). Each ad is coded for its sponsor and the nature of its content, which allows us to determine if an ad is an attack ad or not, a fact we use in constructing some of our covariates. For our outcome data on electoral returns, we used data from CQ's Voting and Elections Collection combined with data on the citizen voting-age population (CVAP) from the US Census. Finally, we obtain polling data from the website RealClearPolitics. We map these county-level outcomes to media markets using the mapping provided by Sides et al. (2022).

Our primary treatment of interest is the presence or absence of independent group (IG) ads. We define D_{it} to be a binary indicator if an IG ran ads in media market i in week t of the campaign. Independent groups are any interest or advocacy group other than the candidate or political party and include so-called 'dark money' groups in addition to political action committees. Let $\overline{D}_{it} = \{D_{i1}, \ldots, D_{it}\}$ be the treatment history up to time t and $\underline{D}_{it} = \{D_{it}, \ldots, D_{iT}\}$ be the history from t

to T. Let $\overline{D}_i \equiv \overline{D}_{iT}$, where these take values in $\mathcal{D}_T \in \{0, 1\}^T$. We investigate several outcomes, including the share of the two-party vote for the Democratic candidate and the share of the eligible vote won by the Democrat and Republican. The latter two outcomes use the CVAP as a denominator, which allows us to explore the possibility that advertising mobilizes each party differently. We denote these outcomes as Y_i and define the potential outcomes $Y_i(\overline{d})$, where $\overline{d} \in \mathcal{D}_T$, which is the outcome that unit i would have if they had followed treatment history $\overline{d}$. We make the usual consistency assumption that $Y_i = Y_i(\overline{d})$ if $\overline{D}_i = \overline{d}$. This assumption implicitly assumes no interference across media markets.

We also have a number of time-varying confounders, including various measures of past advertising by other groups and other candidates and polling averages on support for the Democratic candidate and percent undecided or backing third-party candidates. We denote the measure of these covariates in week t as X_{it}, and we lag these measures carefully to ensure they are causally prior to D_{it}. We define $\overline{X}_{it}$, $\underline{X}_{it}$, and $\overline{X}_i$ similarly to the treatment history.

3 A review of marginal structural models

The combination of marginal structural models and inverse probability of treatment weighting was developed by Robins (1998a) and has since become an important method across a number of scientific domains. Robins et al. (2000) provides a general introduction to the method. A robust methodological literature has built up around the method, focusing on stabilizing the construction of the weights (Cole & Hernán, 2008; Imai & Ratkovic, 2015; Kallus & Santacatterina, 2021; Xiao et al., 2013), using machine learning methods to make estimation more flexible (Gruber et al., 2015; Muñoz & van der Laan, 2011), or developing doubly robust versions of the approach (Bang & Robins, 2005; Rotnitzky et al., 2012). Our contribution to this literature is to show how these methods may be applied when a researcher suspects there may be time-constant unmeasured confounding.

The MSM methodology is based on a sequential ignorability assumption that treatment at time t is unrelated to the potential outcomes conditional on (some function of) the history of treatment and the time-varying covariates. In particular, there is some vector of time-varying covariates, such that,

$$Y_i(\overline{d}) \perp\!\!\!\perp D_{it} \mid \overline{X}_{it}, \overline{D}_{i,t-1}, \quad \forall\, \overline{d} \in \{0, 1\}^T.$$

This assumption is a time-varying version of a selection-on-observables assumption applied repeatedly to treatment in each period. One drawback of this approach in the social sciences is that units may have differing baseline probabilities of treatment based on traits that are difficult to measure. In the context of advertising, groups may target ads at media markets that have more persuadable voters by some metric unknown to the researcher. This limitation of sequential ignorability is one motivation for developing the fixed-effects approach we introduce below.

We are interested in estimating the causal effect of different treatment histories,

$$\mathbb{E}[Y_i(\overline{d}) - Y_i(\overline{d}')],$$

where $\overline{d}, \overline{d}' \in \{0, 1\}^T$. Unfortunately, when T is even moderately large, the number of possible treatment histories grows and it becomes difficult to estimate any particular contrast. To help mitigate this problem, we focus on a marginal structural model, which is a model for the marginal mean of the potential outcomes as a function of the treatment history

$$\mathbb{E}[Y_i(\overline{d})] = g(\overline{d}; \gamma_0), \tag{1}$$

parameterized as a function of γ. Throughout, we use a zero subscript (γ_0, for example) to indicate the true values of parameters. The dimensionality of $\overline{d}$ grows quickly in T, so even when T is moderate, $g(\cdot)$ will usually impose some parametric restrictions on the response surface. Even if these modelling restrictions are correct, the observed conditional expectation function $\mathbb{E}[Y_i \mid \overline{D}_i = \overline{d}] \neq g(\overline{d}; \gamma_0)$ due to confounding by X_{it}. On the other hand, including the covariates in the conditional expectation will lead to post-treatment bias so that $\mathbb{E}[Y_i \mid \overline{D}_i = \overline{d}, \overline{X}_i] \neq g(\overline{d}; \gamma_0)$. Robins (1999) showed how an inverse

probability of treatment weighting scheme could avoid these two biases. In particular, he showed that a weighted conditional expectation can recover the parameters of the MSM when the weights are proportional to the inverse of the conditional probability of the unit's treatment history given their covariate history. Let $\pi_t(\overline{d}_{t-1}, \overline{x}_t) = \mathbb{P}(D_{it} = 1 \mid \overline{D}_{i,t-1} = \overline{d}_{t-1}, \overline{X}_{it} = \overline{x}_t)$ and let $\pi_{it} = \pi_t(\overline{D}_{i,t-1}, \overline{X}_{it})$. Then, the IPTW weights for our MSM become

$$W_i = \prod_{t=1}^{T} \pi_{it}^{-D_{it}} (1 - \pi_{it})^{-(1-D_{it})} \tag{2}$$

With these weights, Robins (1999) showed that $\mathbb{E}[\mathbf{1}\{\overline{D}_i = \overline{d}\} W_i Y_i] = g(\overline{d}; \gamma_0)$.

In observational studies, the propensity scores used to construct the weights are not usually known to the analyst and so must be estimated. The standard approach to this in the MSM literature is to specify a parametric model for treatment and estimate its parameters via maximum likelihood. Define a parametrization of the propensity score $\pi_t(\overline{x}_t, \overline{d}_t; \beta)$, where we define the true value of this parameter as $\pi_t(\overline{x}_t, \overline{d}_t; \beta_0) = \mathbb{P}(D_{it} = 1 \mid \overline{X}_{it} = \overline{x}_t, \overline{D}_{it} = \overline{d}_t)$. We then define the estimated propensity scores as $\widehat{\pi}_{it} = \pi_t(\overline{X}_{it}, \overline{D}_{it}; \widehat{\beta})$, where $\widehat{\beta}$ is the MLE. These estimated propensity scores can then be used to generate estimated weights, $\widehat{W}_i = \prod_{t=1}^{T} \widehat{\pi}_{it}^{-D_{it}} (1 - \widehat{\pi}_{it})^{-(1-D_{it})}$. With these estimated weights, an IPTW estimator for the MSM can be constructed by solving the empirical version of the following estimating equation for γ,

$$\mathbb{E}\left\{ \widehat{W}_i h(\overline{D}_i)(Y_i - g(\overline{D}_i; \gamma)) \right\} = 0,$$

where $h(\cdot)$ is a researcher-specified $\dim(\gamma) \times 1$ vector of fixed functions of $\overline{d}$. This approach finds the value of γ that makes the MSM residuals approximately uncorrelated with $h(\overline{D}_i)$ in the weighted data, and it simplifies to standard estimation techniques in many cases. For example, when $g(\cdot)$ and $h(\cdot)$ are the identity functions, then this approach reduces to weighted least squares. Robins (1998b) established this procedure as producing a consistent and asymptotically normal estimator for the parameters of the MSM.

The weights in equation (2) can often be unstable when the true or estimated propensity scores are close to one or zero, which can lead to highly variable estimates. A common practice, in this case, is to include a stabilizing numerator that is the marginal probability of the treatment history, $\overline{\pi}_{it} = \mathbb{P}(D_{it} = 1 \mid \overline{D}_{i,t-1})$. In this case, the stabilized weights become

$$\widetilde{W}_i = \prod_{t=1}^{T} \left(\frac{\overline{\pi}_{it}}{\pi_{it}} \right)^{D_{it}} \left(\frac{1 - \overline{\pi}_{it}}{1 - \pi_{it}} \right)^{1-D_{it}}.$$

Another common practice is to trim the weights to additionally guard against unstable causal parameter estimates (Cole & Hernán, 2008), though other propensity score estimation techniques also help with this problem (Imai & Ratkovic, 2015).

4 Fixed-effect propensity score estimators

4.1 Setting and assumptions

We now focus on estimating propensity scores with fixed effects for MSMs when time-constant unmeasured confounding exists. As with the traditional MSM case, we assume that $(Y_i, \overline{D}_i, \overline{X}_i)$ are independent across observations. In order to adjust for unit-specific heterogeneity, we do require restrictions beyond the typical MSM case. First and foremost, we focus on marginal structural models for a treatment history of a fixed length rather than the entire treatment history, which we call *truncated* MSMs. In particular, truncated MSMs focus on modelling only the last k periods of treatment,

$\mathbb{E}[Y_i(\underline{d}_{T-k})] = g(\underline{d}_{T-k}; \gamma)$, where $\underline{d}_{T-k} = (d_{T-k}, \ldots, d_T)$, k is fixed, and the parameter vector γ is of length J. Truncation is important for our asymptotic analysis because it limits the number of propensity scores that need to be included in the weights and thus limits the amount of bias induced by the incidental parameters problem in modelling the propensity scores.

Truncation is a restriction on what quantities of interest can be consistently estimated in this setting, not a substantive assumption about the effect of the treatment before the truncation point. By the usual consistency assumption, we can define these 'shorter' potential outcomes as $Y_i(\underline{d}_{T-k}) \equiv Y_i(\overline{D}_{i,T-k-1}, \underline{d}_{T-k})$, so that treatment history before k lags acts more like a baseline confounder. In particular, our use of a truncated MSM does not invoke a 'no carryover' assumption as in Imai and Kim (2019). Compared to typical MSM practice, the main limitation of this restriction is to rule out functional forms where the cumulative sum of the entire treatment history is included as part of the MSM. Intuitively, this restriction implies that analysts cannot simultaneously use long treatment histories to estimate long-term effects and adjust for unmeasured confounding.

We now describe the key identification assumption of the IPTW-FE approach, which combines the concept of unit-specific randomized experiments with the standard MSM framework in Section 3. Let $\underline{X}_{i,t+1}(\overline{d}) = (X_{i,t+1}(\overline{d}_t), X_{i,t}(\overline{d}_{t+2}), \ldots, X_{i,T}(\overline{d}_{T-1}))$ represent the potential outcomes of the future covariates under a particular treatment history, where we truncate the full treatment history $\overline{d} = (d_1, \ldots, d_T)$ for each time period, $\overline{d}_k = (d_1, \ldots, d_k)$, since future treatment values cannot affect past covariates. This latter property is also known as a *no anticipation* assumption.

Assumption 1 (Unit-specific Sequential Ignorability). Let α_i be an unmeasured, time-constant random variable. For all i, t and $\overline{d}$,

$$\{Y_i(\overline{d}), \underline{X}_{i,t+1}(\overline{d})\} \perp\!\!\!\perp D_{it} \mid \overline{X}_{it}, \overline{D}_{i,t-1} = \overline{d}_{t-1}, \alpha_i.$$

Assumption 1 states that conditional on the unit-specific effect, the treatment history, and (a function of) the covariate history, treatment is independent of future potential outcomes for both the outcome and the covariate process. In essence, treatment is randomized with respect to future covariates and the outcome, conditional on the past and time-constant features of the unit. This assumption allows for both time-varying confounding by measured covariates and time-constant confounding by measured and unmeasured covariates. We do assume that the time-constant unmeasured confounding can be captured by the unidimensional value α_i, which might represent a combination of several unit-specific factors. This assumption will be violated (and the theoretical results about our estimator invalid) if there is time-varying unmeasured confounding, a feature our approach shares with most panel data estimators for causal inference such as linear fixed effects and difference-in-differences.

Assumption 1 involves potential outcomes of the entire treatment history, $Y_i(\overline{d})$, but above, we defined our main marginal structural models in terms of truncated treatment histories, $\mathbb{E}[Y_i(\underline{d}_{T-k})]$. Thus, the requirements of sequential ignorability go beyond the treatments of interest in the marginal structural model and apply to the potential outcomes for the entire treatment history. This allows for the fixed-effect propensity score estimators to be consistent even without a no-carryover assumption that would assume that treatment before $T-k$ has no effect on the outcome.

Under Assumption 1, we can nonparametrically identify the mean of the potential outcomes under a given history with unit-specific propensity scores. Let $\pi_{it}(\overline{x}_t, \overline{d}_{t-1}, \alpha_i) = \mathbb{P}(D_{it} = 1 \mid \overline{X}_{it} = \overline{x}_t, \overline{D}_{i,t-1} = \overline{d}_{t-1}, \alpha_i)$ and let $\pi_{it} = \pi_{it}(\overline{X}_{it}, \overline{D}_{i,t-1}, \alpha_i)$. Then, we can use the usual techniques to arrive at the nonparametric identification of

$$\mathbb{E}[Y_i(\underline{d}_{T-k})] = \mathbb{E}\left[\frac{\mathbf{1}(\underline{D}_{i,T-k} = \underline{d}_{T-k})Y_i}{\prod_{t=T-k}^{T} \pi_{it}^{d_t}(1-\pi_{it})^{1-d_t}}\right],$$

where d_t denotes the corresponding entry in $\underline{d}_{T-k}$. Thus, under Assumption 1 (and a positivity assumption), treatment history effects are nonparametrically identified since we can write them as functions of quantities that are in principle observable as $N, T \to \infty$.

As is common with nonparametric identification, however, the sampling details across units and time will play an important role in actually obtaining valid estimates of these causal effects. We

can see this even in static causal inference settings. If units in that setting are not independent across units, for example, standard IPW approaches might not be estimable at standard rates without further assumptions. While we assume i.i.d. data across units, this assumption would be unrealistic for the time dimensions. We now lay out the sampling assumptions for our setting.

Assumption 2 (Sampling Assumptions).

(i) (Asymptotics) Let $N, T \to \infty$ such that $N/T \to \rho$ where $0 < \rho < \infty$.

(ii) (Across/Within-Unit Dependence) For all N and T, $\{(Y_i(\bullet), \overline{D}_i, \overline{X}_i, \alpha_i) : i = 1, \ldots, N\}$ are i.i.d. across i, where $Y_i(\bullet) = \{Y_i(\overline{d}); \overline{d} \in \{0,1\}^T\}$. Letting $Z_{it} = (D_{it}, X_{it})$ for $t = 1, \ldots, T$ and $Z_{i,T+1} = (Y_i(\overline{d}))$, then for each i, $\{Z_{it} : t = 1, \ldots, T+1\}$ is α-mixing conditional on α_i with mixing coefficients satisfying $\sup_i a_i(m) = O(m^{-\mu})$ as $m \to \infty$ where $\mu > 4(8+\nu)/\nu$ and $\nu > 0$,

$$a_i(m) \equiv \sup_t \sup_{A \in \mathcal{A}_{it}, B \in \mathcal{B}_{i,t+m}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

$\mathcal{A}_{it}$ is the sigma field generated by $(Z_{it}, Z_{i,t-1}, \ldots)$, and $\mathcal{B}_{i,t}$ is the sigma field generated by $(Z_{it}, Z_{i,t+1}, \ldots)$.

Assumption 2(i) establishes the large-N, large-T asymptotic framework, which has been widely used for nonlinear panel models in econometrics (Arellano & Hahn, 2007; Fernández-Val, 2009; Fernández-Val & Weidner, 2016, 2018; Hahn & Kuersteiner, 2011; Hahn & Newey, 2004). The strong mixing process in Assumption 2(ii) allows us to rely on the laws of large numbers and the central limit theorem in the time dimension. It essentially states that dependence over time is sufficiently weak that as the distance between two periods increases, information in the two periods becomes approximately uncorrelated. That is, data over time within a unit may be dependent, but there is new information as time goes on. This assumption is substantially weaker than independence over time or even stationarity. In particular, it allows for time trends, which are a common feature of propensity score models in MSMs. The i.i.d. nature of the distribution of the data and the fixed effects across units is common to IPTW approaches and allows us to take averages over the unit-specific heterogeneity and has been used before for average partial effects in nonlinear panel models (Fernández-Val & Weidner, 2016). It is possible to replace this assumption with stationarity of X_{it} over time, but this would rule out lagged treatment in the propensity score model along with time trends.

To determine the asymptotic properties of our approach, we assume researchers will specify a correct parametric model for the propensity score (up to the unmeasured heterogeneity) as $\pi_{it}(\overline{x}_t, \overline{d}_{t-1}; \beta, \alpha_i) = \mathbb{P}(D_{it} = 1 \mid \overline{X}_{it} = \overline{x}_t, \overline{D}_{i,t-1} = \overline{d}_{t-1}; \beta, \alpha_i)$, where β is a $k \times 1$ parameter vector, α_i is the time-constant unmeasured confounder, and $\pi_{it}(\beta, \alpha_i) = \pi_{it}(\overline{X}_{it}, \overline{D}_{i,t-1}; \beta, \alpha_i)$. We write the log-likelihood of this model as

$$\ell_{it}(\beta, \alpha) = D_{it} \log \pi_{it}(\beta, \alpha) + (1 - D_{it}) \log\{1 - \pi_{it}(\beta, \alpha)\}$$

Let $\alpha_0 = (\alpha_{10}, \ldots, \alpha_{N0})$ and β_0 be the values of the parameters that generate the treatment process. In particular, we assume that these values are the solution to the following population conditional maximum likelihood condition

$$(\beta_0, \alpha_0) = \arg\max_{(\beta,\alpha) \in \mathbb{R}^{d_\beta + N}} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}[\ell_{it}(\beta, \alpha) \mid \alpha_i], \tag{3}$$

where the expectation is with respect to the distribution of the data conditional on the unobserved effect (see, for example, Fernández-Val & Weidner, 2016, equation 2.1). Our theoretical results require a correctly specified parametric model for the covariates in the propensity score (which is common in the MSM literature), but the approach is semiparametric in that we make no assumptions about the

relationship between the unmeasured heterogeneity and the covariates. In simulation studies in the Supplemental Materials, we also find that the approach is also robust to certain forms of misspecification of the propensity score (namely, the link function). We assume a fixed-length parameter vector, but it may be possible to allow this vector to grow with N and T and thus allow for more flexible estimation strategies. As this is beyond the scope of the current paper, we leave it to future research.

With this propensity score model in hand, we can construct weights that can adjust for both observed time-varying confounding and unobserved time-constant confounding. In particular, we use the following weights

$$W_i(\beta, \alpha_i) = \prod_{j=T-K}^{T} \left(\frac{1}{\pi_{ij}(\beta, \alpha_i)}\right)^{D_{ij}} \left(\frac{1}{1-\pi_{ij}(\beta, \alpha_i)}\right)^{1-D_{ij}},$$

where we only take the product over the last k time periods because our quantities of interest focus on those periods. As with the standard MSM case, we can replace the numerator with the marginal probability of the treatment history, $\overline{\pi}_{it}$, which can stabilize the variance of the estimator without affecting identification.

The IPTW approach to estimating this MSM is to rely on the estimating equation

$$0 = \frac{1}{N}\sum_{i=1}^{N} U_i(\gamma, \beta, \alpha_i) = \frac{1}{N}\sum_{i=1}^{N} \left\{ W_i(\beta, \alpha_i) h(\underline{D}_{i,T-k})(Y_i - g(\underline{D}_{i,T-k}; \gamma)) \right\},$$

where $h(\cdot)$ is a function with J-length output chosen by the researcher as in the standard MSM case. For example, if Y_i is continuous and g is linear and additive, it is common to use $h(\underline{D}_{i,T-k}) = \underline{D}'_{i,T-k}$. Under the fixed-effects sequential ignorability assumption and the MSM, we have $\mathbb{E}[U_i(\gamma_0, \beta_0, \alpha_{i0})] = 0$, which is a semiparametric identification result because the restriction identifies the causal parameters, γ_0, solely in terms of sample quantities (up to the propensity score parameters). This result follows the standard g-computation algorithm with the unit-specific heterogeneity, α_i, included in the place of a baseline covariate (Robins, 1999, 2000).

4.2 Proposed method

We propose a two-step approach to estimating the parameters of the marginal structural model using inverse probability of treatment weighting. These two steps are:

1. Obtain estimates of the parameters of the propensity score model, $(\widehat{\beta}, \widehat{\alpha}_i)$, using conditional maximum likelihood treating the unit-specific effects α_i as fixed parameters to be estimated. Construct estimated weights $W_i(\underline{D}_{i,T-k}; \widehat{\beta}, \widehat{\alpha}_i)$.
2. Pass the estimated weights to a weighted estimating equation $N^{-1}\sum_{i=1}^{N} U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i) = 0$ to obtain estimates of the MSM parameters, γ.

The first step in this procedure can be implemented with a sample conditional maximum likelihood estimator. Letting $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_N)$, we have

$$(\widehat{\beta}, \widehat{\alpha}) = \underset{(\beta,\alpha)\in\mathbb{R}^{d_\beta+N}}{\arg\max} \sum_{i=1}^{N}\sum_{t=1}^{T} \ell_{it}(\beta, \alpha_i) \tag{4}$$

Under these assumptions, we use the following maximum likelihood estimators:

$$\widehat{\beta} = \underset{\beta}{\arg\max} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \ell_{it}(\beta, \widehat{\alpha}_i(\beta)), \qquad \widehat{\alpha}_i(\beta) = \underset{\alpha}{\arg\max} \frac{1}{T}\sum_{t=1}^{T} \ell_{it}(\beta, \alpha).$$

These maximum likelihood estimates are subject to the usual incidental parameters problem that results in bias that shrinks as $T \to \infty$. Even when N and T grow at the same rate, Hahn and Newey (2004) showed that these types of MLE estimators are not $\sqrt{NT}$-consistent, and a large literature has developed proposing several bias correction techniques (Arellano & Hahn, 2007; Fernández-Val & Weidner, 2018). We sidestep these issues in our results because we target the slower convergence rate of $\sqrt{N}$, since we only have a single outcome per unit, which is common in the MSM literature.

To obtain estimates of the MSM parameters, $\widehat{\gamma}$, we use the sample version of the MSM moment condition, $N^{-1}\sum_{i=1}^{N} U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i) = 0$. This estimator depends on the link function for the marginal structural model and a function $h(\cdot)$. One particularly straightforward estimator in this class is weighted least squares for the identity link with continuous outcomes. Often, $h(\cdot)$ can be chosen to enhance the efficiency of the estimator (Robins, 1999), but we do not explore that here. We now show in Theorem 1 that under the above assumptions and some regularity conditions, this estimator is consistent and asymptotically normal. The proof and precise statements of the regularity conditions are in online supplementary material, Supplemental Materials A. Let $G = \mathbb{E}\{\partial U_i(\gamma, \beta, \alpha)/\partial\gamma\}_{\gamma=\gamma_0}$, and $U_i = U_i(\gamma_0, \beta_0, \alpha_{i0})$.

Theorem 1 Under Assumptions 1, 2, and suitable regularity conditions, $\widehat{\gamma} \xrightarrow{p} \gamma_0$ and

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{d} N(0, V_{\gamma_0}), \tag{5}$$

where $V_{\gamma_0} = G^{-1}\mathbb{E}[U_i U_i^{\top}]G^{-1}$.

We can build a consistent variance estimator in the usual way with $\widehat{V}_\gamma = \widehat{G}^{-1}\widehat{\Omega}\widehat{G}^{-1}$, where

$$\widehat{G} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \widehat{U}_i}{\partial \gamma}, \quad \widehat{\Omega} = \frac{1}{N}\sum_{i=1}^{N}\widehat{U}_i\widehat{U}_i^{\top}, \quad \widehat{U}_i = U_i(\widehat{\gamma}, \widehat{\beta}, \widehat{\alpha}_i).$$

This is a standard sandwich estimator for estimators based on estimating equations.

Theorem 1 establishes that the IPTW-FE for MSMs is asymptotically normal and that we can asymptotically ignore the estimation of the weights. In the standard IPTW case, the estimation of the weights does impact the distribution of the MSM estimates. Here, however, the estimation of the weights doesn't affect the first-order asymptotic distribution because we are using NT observations to estimate the propensity score parameters but only using a fraction of the observations, Nk, to create the weights, where k is fixed as $T \to \infty$. Thus, the $\widehat{\beta}$ converges much faster than $\widehat{\gamma}$ and so we can ignore its estimation error. Of course, this is an approximation that might be less accurate when T is small, so a bootstrap of units might yield more accurate variance estimates in that case. (However, in a simulation study, we found that the nonparametric bootstrap has almost exactly the same performance as this estimator.) Under a specific model for the propensity score, one could also derive the second-order impact of the propensity score model and derive an analytical expression for the variance.

In typical nonlinear panel models, plugging in noisy estimates of the fixed-effect parameters leads to a bias that converges to 0 slowly enough to cause, for example, $\sqrt{NT}(\widehat{\beta} - \beta_0)$ to not be asymptotically centered at 0. In our setting, however, the strong mixing property of the treatment process ensures that this bias fades over time, and so allows us to ignore the estimation of the fixed-effect parameters as well. In the literature on nonlinear panel models, there is a similar result for estimating partial effects or differences in the conditional expectation, as opposed to parameters of the nonlinear model. For example, Fernández-Val and Weidner (2018) showed how these average partial effects can converge at a slower rate with parameter estimation not having a first-order effect on the asymptotic distribution (see also Fernández-Val & Weidner, 2016). The current approach is similar since we are only interested in the parameters of the weighting model insofar as they provide consistent estimates of the IPTW weights.

This result establishes that it is possible to adjust for unmeasured baseline confounding in MSMs when the time dimension is long and provides sufficiently new information within units. The quality of this adjustment will depend on both how long the panels are and how severe the unmeasured

heterogeneity is. A second-order expansion of the estimator shows that second-order bias (which can be ignored in our asymptotic analysis) is inversely related to the propensity scores. Thus, strong unit-specific heterogeneity that pushes propensity scores close to zero or one could create more finite-sample bias, though this bias will be second order compared to the first-order asymptotic bias of ignoring the unmeasured confounding. Indeed, as we find in our simulations, our estimator performs worse under strong (versus weak) unobserved heterogeneity, but it improves over a naive IPTW approach that ignores the heterogeneity under either scenario. Longer panels help with this finite-sample bias since these second-order terms will be of order $O_P(1/\sqrt{T})$. A fruitful avenue for future research would be to use analytic or computational approaches like the jackknife to adjust for these second-order terms as in Hahn and Newey (2004).

What about doubly robust estimation? In traditional MSM settings, it is possible to develop doubly robust estimators that depend both on the correct modelling of the propensity scores and a series of outcome regression models (Bang & Robins, 2005). In our setting, however, this would require an outcome regression model that had unobserved heterogeneity, and without multiple observations of the outcome over time, it is not possible to estimate such a model without overly strong assumptions.

4.3 Trimming weights

One drawback of the IPTW-FE approach is that the fixed-effect parameters of the propensity score model are not identified when units are either always treated or always control. Even when we maintain the population-level positivity assumption, this in-sample positivity violation means that some units will have undefined weights. We propose three ways to address this issue. First, one could simply omit the no-treatment-variance units and estimate the parameters of the MSM for the units that have at least one treated and one control period. This is the simplest procedure but could induce confounding bias, especially if the α_i has a nonlinear relationship with the outcome. Second, we could use an ad hoc rule for imputing propensity scores of the no-treatment-variance units. For example, we could set these units to have $\widehat{\pi}_{it} = 0.01$ if $D_{it} = 0$ for all t and $\widehat{\pi}_{it} = 0.99$ if $D_{it} = 1$ for all t. Depending on the lag length k in the MSM and the exact trimming, this may lead to extreme weights, which themselves could require trimming. Alternatively, one could place bounds on the range of the unit-specific effects in the MLE estimation to $\alpha_i \in [a_0, a_1]$ and set the estimates of those effects as $\widehat{\alpha}_i = a_0$ or $\widehat{\alpha}_i = a_1$ if $D_{it} = 0$ or $D_{it} = 1$ for all t, respectively. The amount of trimming of the weights in this approach amounts to a bias-variance trade-off similar to weight trimming in standard IPTW estimators for MSMs (Cole & Hernán, 2008).

Finally, one alternative approach to handling positivity violations would be to focus on a different quantity of interest. Kennedy (2019) proposed estimating the effect of incremental propensity score interventions, which are interventions that shift the propensity score rather than set treatment histories to specific values. The identification and estimation of these effects do not depend on positivity, and under the assumption of a correctly specified propensity score model, a simple inverse probability weighting estimator is available (Kennedy, 2019, p. 650).

5 Simulation evidence

In this section, we conduct simulation studies to evaluate the finite sample performance of the proposed approach.

5.1 Setup

We simulate a balanced panel of n units with T time points where the number of units varies $n \in \{200, 500, 1{,}000, 3{,}000\}$. We fix the ratio of the number of units to the number of time periods $n/T = \rho \in \{5, 10, 50\}$. This setup mimics the key asymptotic approach of our theoretical results, and the larger value of ρ implies a small number of time points, $T = n/\rho$. The treatment sequence is

generated as a function of the individual unobserved effect α_i, the past treatment $D_{i,t-1}$ and the time-varying covariates, X_{it}.

$$D_{it} \sim \text{Bernoulli}(\text{expit}(\alpha_i + \varphi D_{i,t-1} + \beta^\top X_{it}))$$

where $\text{expit}(x) = 1/(1+\exp(-x))$ is the inverse logistic function. The individual heterogeneity is drawn from a uniform distribution with support on $[-a, a]$ for $a \in \{1, 2\}$. The value of a is chosen such that the variance of individual heterogeneity explains $1/3$ ($a = 1$) or $2/3$ ($a = 2$) of the variance of the linear predictor. The time-varying covariates X_{it} are generated exogenous to the treatment, drawn from the multivariate normal distribution, $X_{it} \sim \mathcal{N}(-1/2\mathbf{1}, \Sigma)$ where $\Sigma_{jj} = 1$ and $\Sigma_{jj'} = 0.2$ for $j \neq j'$. Finally, we set $\varphi = 0.3$ and $\beta = (-0.5, -0.5)$ when the number of covariates is two or $\beta = (-0.5, -0.5, 1.0, -0.5)$ when the number of covariates is four.

The outcome is generated by the linear model with individual unobserved variable α_i, the final treatment D_{iT}, the cumulative lagged treatments $\sum_{t=T-3}^{T-1} D_{it}$ and the average of the time-varying covariates, $\overline{X}_i = \sum_{t=1}^{T} X_{it}/T$, all of which are generated in the previous step.

$$Y_i = \alpha_i + \tau_F D_{iT} + \tau_C \sum_{t=T-3}^{T-1} D_{it} + \gamma^\top \overline{X}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

where we set $\tau_F = 1$, $\tau_C = 0.3$, and $\gamma = (1.0, 0.5)$ or $\gamma = (1.0, 0.5, 1.0, 1.0)$ depending on the number of covariates used in each simulation.

5.2 Results

We compared the performance of the proposed method in terms of estimating two causal quantities: the final period effect τ_F and the cumulative lagged effect τ_C. We estimate two quantities together in the framework of weighted least squares,

$$(\widehat{\tau}_F, \widehat{\tau}_C) = \arg\min_{\tau_F, \tau_C} \sum_{i=1}^{n} \widehat{W}_i \left\{ Y_i - \alpha - \tau_F D_{iT} - \tau_C \sum_{t=T-3}^{T-1} D_{it} \right\}^2$$

where $\widehat{W}_i$ is constructed as described in the previous section. We focus on this correctly specified MSM to isolate the effects of unmeasured heterogeneity on estimator performance. The variance of $\widehat{\tau}_F$ and $\widehat{\tau}_C$ is estimated using the standard sandwich formula with the HC2 option, which is an adjustment to improve finite-sample properties of the variance estimator (MacKinnon & White, 1985).

In addition to the fixed effect approach, we consider two other strategies to obtain the weights $\widehat{W}_i$ as benchmarks to the proposed method. First, we use the true propensity score to construct the weights. Second, the estimated propensity score without the fixed effect is used to construct weights. We expect that the weights with known propensity scores are least biased and the weights without the fixed effect are most biased.

Figure 1 shows the results for the two-covariate case. Bias (first row), standard errors (second row) and coverages (third row) are computed based on 500 Monte Carlo simulations. Additional simulation results are presented in online supplementary material, Supplemental Materials C. The first two columns correspond to the 'low' heterogeneity case where the support of the fixed effect is $[-1, 1]$, whereas the last two columns correspond to the 'high' heterogeneity scenario where the support of α_i is set to $[-2, 2]$. Solid lines in blue show the proposed estimator (`IPTW-FE`), solid lines in grey show the estimator based on the true propensity score (`IPTW-True`), and dashed lines in green show the estimator based on the estimated propensity score without fixed effects (`IPTW`). Shapes correspond to the n to T ratio ρ such that squares represent $\rho = 5$ (the largest number of time periods), circles represent $\rho = 10$, and triangles represent $\rho = 50$ (the smallest number of time periods).

We can see that under the low heterogeneity setting, where the unobserved individual heterogeneity explains roughly $1/3$ of the variance of the treatment assignment, the bias of the proposed estimator (`IPTW-FE`) is indistinguishable from the estimator that is based on the true propensity score

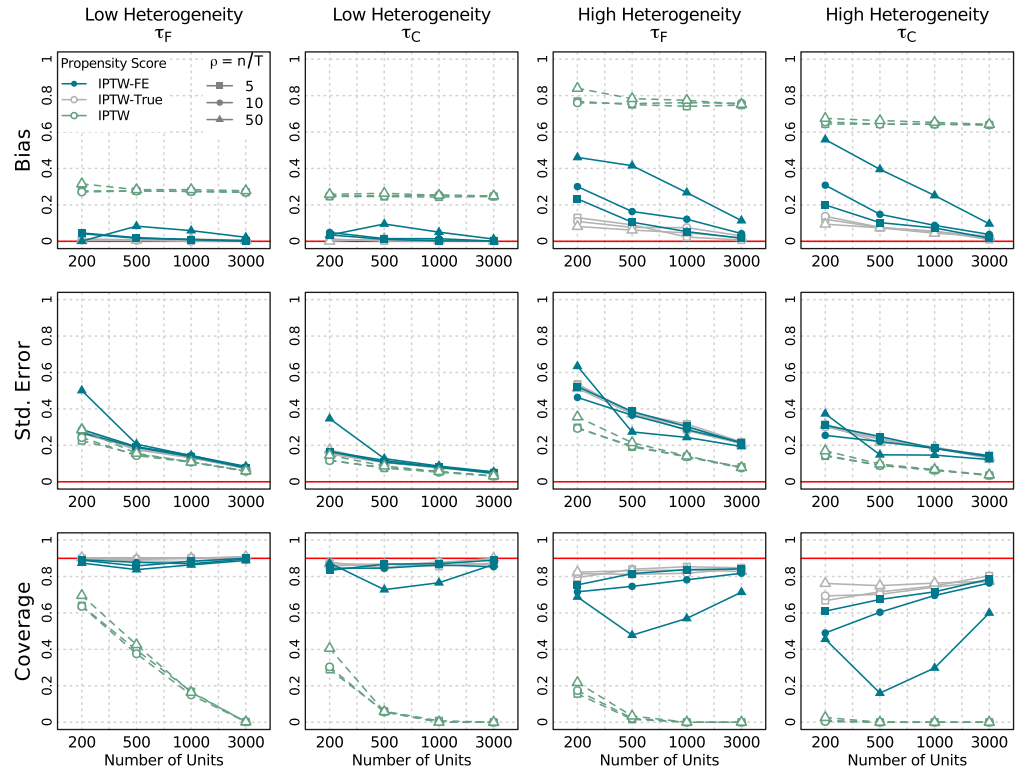

Figure 1. Bias, standard error (Std. Error) and coverage probability of 90% confidence intervals (Coverage) for the estimation of the final period effect τ_F and the cumulative effect τ_C under the 'low' heterogeneity ($a = 1$)—first two columns—and the 'high' heterogeneity ($a = 2$)—last two columns—scenario. Solid lines in blue show the proposed estimator (`IPTW-FE`), solid lines in grey show the estimator based on the true propensity score (`IPTW-True`), and dashed lines in green show the estimator based on the estimated propensity score without fixed effects (`IPTW`). Shapes correspond to the n to T ratio ρ such that squares represent $\rho = 5$ (the largest number of time periods), circles represent $\rho = 10$, and triangles represent $\rho = 50$ (the smallest number of time periods).

(`IPTW-True`) and the confidence interval estimates maintain the nominal coverage across different values of n and ρ. Under this scenario, even in the case of $n = 200$ and $T = 4$, the proposed method performs well. The naive `IPTW` approach exhibits higher bias in this setting but also lower sampling uncertainty than `IPTW-FE`, highlighting a trade-off between lowering bias and inflating standard errors. If unobserved heterogeneity was very low or nonexistent, then `IPTW` might provide better finite-sample performance, especially in small samples.

When the variance of the individual heterogeneity is high ($a = 2$) such that it explains roughly 2/3 of the variance of the treatment assignments, the proposed estimator shows relatively larger bias compared with `IPTW-True`, while bias of the estimator without fixed effects (`IPTW`) is substantially larger. Furthermore, the bias of our approach shrinks as the number of units and periods grows (and our estimates of the propensity score improve), whereas the bias of the naive `IPTW` approach stays constant as n and T grow. Under this setting, the coverage results are mainly driven by the bias, thus the figure shows that as n increases, the coverage results also improve thanks to the reduction in bias. We can also see that in general the estimator without fixed effects shows smaller standard errors than `IPTW-FE`, again highlighting the bias-variance trade-off. This implies that the proposed method (`IPTW-FE`) trades off efficiency for lower bias. Finally, we highlight that small Monte Carlo bias is observed even for `IPTW-True` under this scenario. This is possibly due to the high variability of the weights, which are a product of inverse probabilities over four time periods with stabilization.

Overall, these results point to two key tensions in controlling for time-constant unmeasured heterogeneity through fixed effects in the propensity score models. First, high degrees of unmeasured heterogeneity in the propensity scores may lead to near violations of the positivity assumption that could lead to the kind of instability we see when $a = 2$. Second, larger magnitudes of heterogeneity may require more time periods to achieve good finite sample performance compared to when the heterogeneity is relatively small. As shown in these results, though, the proposed approach can outperform the naive approach in spite of these issues.

In online supplementary material, Supplemental Materials C, we provide additional simulation results when we apply `IPTW-FE` with a misspecified link function in the propensity score. Those results demonstrate that the proposed estimator is fairly robust to this type of misspecification in terms of bias and coverage.

6 Results

6.1 Specification and balance

We now apply these techniques to estimate the effectiveness of independent group advertising in US Senate and Gubernatorial elections from 2010 until 2020. We build on Blackwell (2013), who investigated the effects of negative advertising using an MSM approach without fixed effects for elections over the period from 2000 to 2008. Our primary results focus on three outcomes: the Democratic percentage of the two-party vote, percent of the voting-eligible population casting Democratic votes (which we call 'Democratic Turnout'), and percentage of the voting-eligible population casting Republican votes (which we call 'Republican Turnout').

To calculate the propensity scores, we organize the data into a market-race-week panel, where an example of a market-race would be the 2010 California Gubernatorial election in the Santa Barbara media market (as distinct from the media markets of San Diego, Fresno, and so on). We focus on the time period between the primary election for the race and the general election so that we have campaign lengths ranging from 8 to 40 weeks with a median of 20 weeks. After dropping market-races that have no variation in the treatment, we have $N = 467$ market-races for Democratic IG ads and $N = 623$ market-races for Republican IG ads. Most of the dropped races are very uncompetitive races or media markets with smaller audiences. In online supplementary material, Supplemental Materials D, we investigate an alternative approach to handling no-treatment-variation market-races that uses extreme values of the unit fixed effects to obtain weights. We find results from this approach are very similar to our own below.

Table 1 shows the distribution of our aggregated treatment variable across different election years. Even with only a handful of time periods and a high level of aggregation, we can see that empirical positivity violations in the final weeks of the race are fairly common, which is the main motivation for using a marginal structural model. Note that even if a market-race had zero weeks of ads in the final five weeks of the campaign, it may have had IG ads earlier in the campaign, allowing us to estimate propensity scores for these units. Finally, we note that our theoretical results rely on stability in

Table 1. Number of weeks with democratic IG ads in the last 5 weeks of the campaign.

# of ads weeks	2010	2012	2014	2016	2018	2020	All
0	12	14	25	6	5	6	68
1	10	9	13	5	16	3	56
2	19	8	17	8	14	15	81
3	14	4	6	7	4	8	43
4	13	15	7	4	6	6	51
5	11	34	28	21	50	24	168

the propensity score model over the weeks of the campaign and can accommodate changes in the propensity score model across election years.

Our marginal structural model is

$$\mathbb{E}[Y_i(\bar{d}) \mid R_i] = \gamma_{R_i} + \gamma_1 \left(\sum_{k=0}^{4} d_{T-k} \right),$$

where the time index here is weeks of the campaign and R_i is the electoral race associated with market-race i. This MSM allows for race-specific intercepts, which helps to purge any remaining race-specific confounding from our estimates. The main quantity of interest, γ_1, can be interpreted as the effect of an additional week of IG advertising in the last five weeks on the outcomes, conditional on the state-wide race. For the outcome MSM, we restrict our attention to races with multiple markets to accommodate the race-specific intercepts.

We apply several different estimation approaches to this MSM: the proposed IPTW-FE approach, a standard IPTW approach without fixed effects, and a naive approach that ignores time-varying covariates altogether. For the weighting model, we included various time-varying covariates: average Democratic share of the two-party preferences in polls in the previous week (and the square of this term), the average percentage reporting undecided or voting for third-party candidates in the previous week, measures of Republican negativity over the last six weeks, the cumulative number of ads shown by the Democrat and Republican (and their squared terms). For the fixed effects approach, we additionally include a market-race fixed effect term in the specification. For the IPTW approach, we only include fixed effects at the race level.

The key assumption of these weighting methods is that balancing by these weights is sufficient to adjust for any unmeasured confounding. For the IPTW-FE approach, this requires the relationship between time-varying covariates and treatment to be relatively stable between 2010 and 2020, conditional on the particular market-race. To test this assumption, in online supplementary material, Figure SM.5 in the Supplemental Materials we replicate our main findings with a propensity score model that interacts all covariates with a linear time trend, and there are no major differences from our main results below.

Assessing balance with the IPTW-FE approach is difficult because we care about the balance with respect to both measured and unmeasured confounders. Of course, we cannot assess balance with respect to unmeasured confounders. We can, however, investigate how well IPTW-FE balances the measured time-varying covariates. To do so, we regress each of these covariates on the treatment indicator, the lagged cumulative sum of treatment, and a race-specific intercept (all variables included in the MSM and the numerator of our weighting models) in the five-week period of our MSM. Note that because this includes treatment over time, this procedure checks the quality of the weights at each point in time. Figure 2 shows the distribution of standardized partial correlations of the treatment indicator and the various covariates under different weighting schemes (no weighting, IPTW, and IPTW-FE). Both IPTW and IPTW-FE vastly reduce the conditional imbalance on these covariates relative to the naive approach. In the unweighted approach, there are a few extremely unbalanced time-varying confounders.

6.2 Main results

Figure 3 shows the results of these methods for each of the outcomes. Substantively, the methods generally agree that there is a positive effect of Democratic IG ads on Democratic electoral performance. This effect is driven by a positive effect on Democratic turnout and a weaker negative effect on Republican turnout. Thus, it appears that Democratic independent group ads mobilize Democratic voters and perhaps demobilize Republican voters. Republican IG ads, on the other hand, have no estimated effect on any of these measures, indicating that these ads are not very effective. The different methods here generally agree on the direction and significance of the effects, though IPTW-FE estimates a larger effect for Democratic groups than the basic IPTW approach. For instance, the

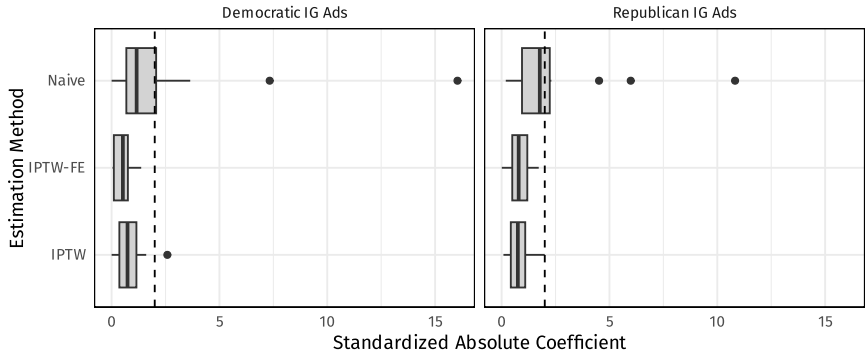

Figure 2. Balance of baseline covariates under different weighting approaches.

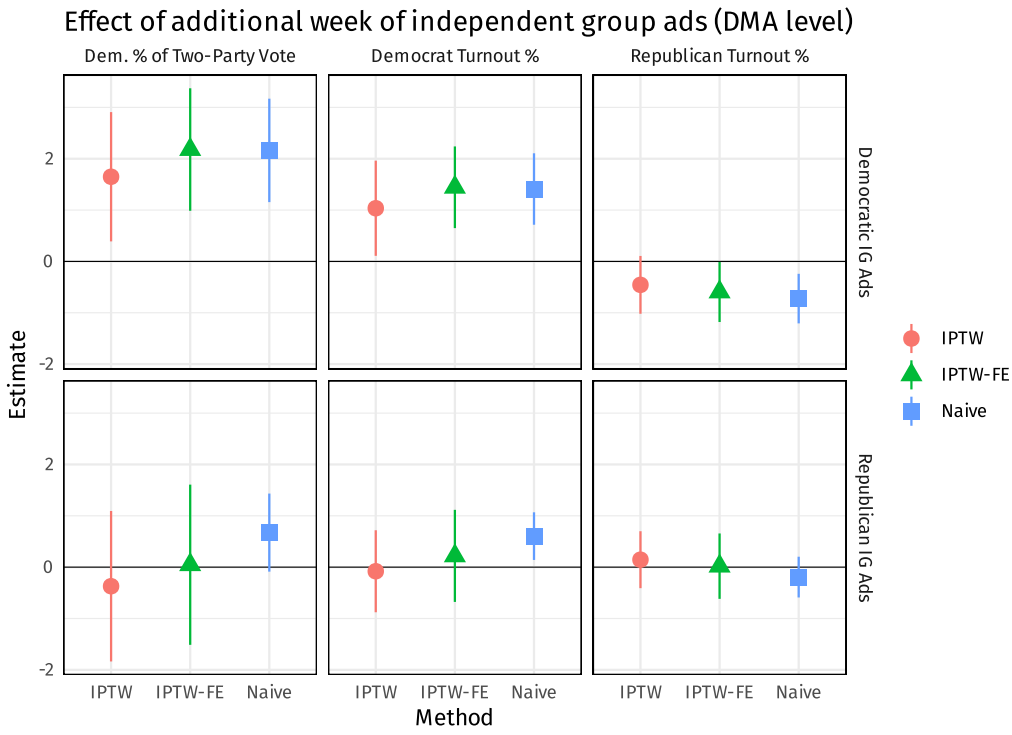

Figure 3. Estimated effects of the number of weeks of independent group advertising in the last five weeks of the campaign with different methods.

IPTW-FE estimate of the effect of Democratic advertising on Democratic vote share is 32% larger than the standard IPTW estimate (2.18 vs. 1.65 percentage points per additional advertising week), and the IPTW-FE estimate on Democratic turnout is 40% larger (1.44 vs. 1.04). These differences are not statistically significant, as the study is not well-powered to detect differences of this magnitude given the relatively small number of races. Nevertheless, the consistent direction across outcomes suggests that the market-race fixed effects in the IPTW-FE propensity score capture confounding not addressed by the standard approach.

The effectiveness of Democratic independent group ads runs counter to the conventional wisdom about what party would benefit the most from the *Citizens United* decision. To understand what drives

this effect, we estimated differential effects by election era. Specifically, we included an interaction between our cumulative treatment measure and an indicator for whether the election was before or after Donald Trump became a candidate for president in 2015. Figure 4 shows that the effectiveness of Democratic group ads is driven in large part by the post-Trump era. These ads are more effective at increasing Democratic votes and reducing Republican votes, and all of these effect differences are statistically significant at the $\alpha = 0.1$ level. In particular, the demobilizing effect of Democratic group ads on Republican voters is a feature of the Trump era. These patterns are consistent with how Trump alienated large segments of Republicans and perhaps made them more vulnerable to ads that encouraged them to stay home or vote for Democrats.

The flexible structure of marginal structural models allows us to investigate which weeks of the campaign are driving the effects on these outcomes. To do so, we can break up the cumulative sum of treated weeks into the number of treated weeks within three weeks of election day and the number of treated weeks 4–5 weeks before election day. Under our assumptions, this is another valid way to parameterize the MSM, allowing us to summarize the causal response surface in a different way. Figure 5 shows the estimated effects of independent group ads at various weeks before election day, as estimated by the IPTW-FE with the baseline covariates. The only major difference is that the positive effect of Democratic IG ads appears stronger in the last few weeks of the campaign compared to earlier weeks. This increased effectiveness of more recent ads is consistent with previous experimental studies of television ads (Gerber et al., 2011).

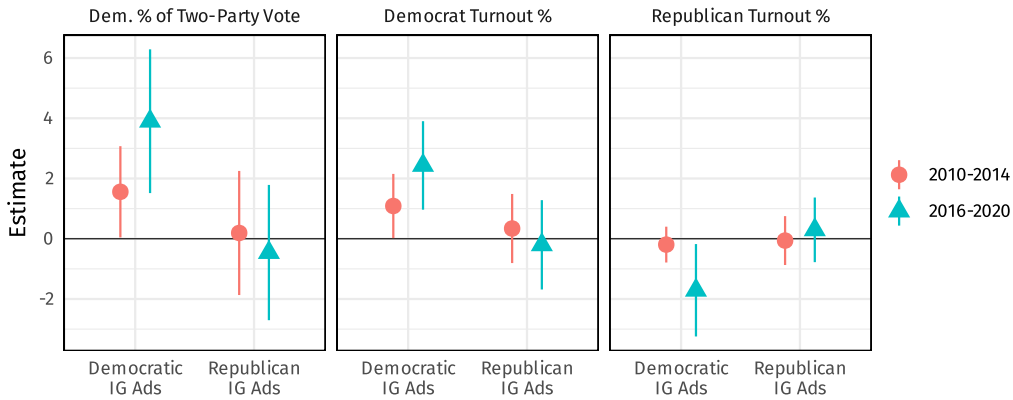

Figure 4. Treatment effect heterogeneity before and after Donald Trump enters the 2016 Presidential Race.

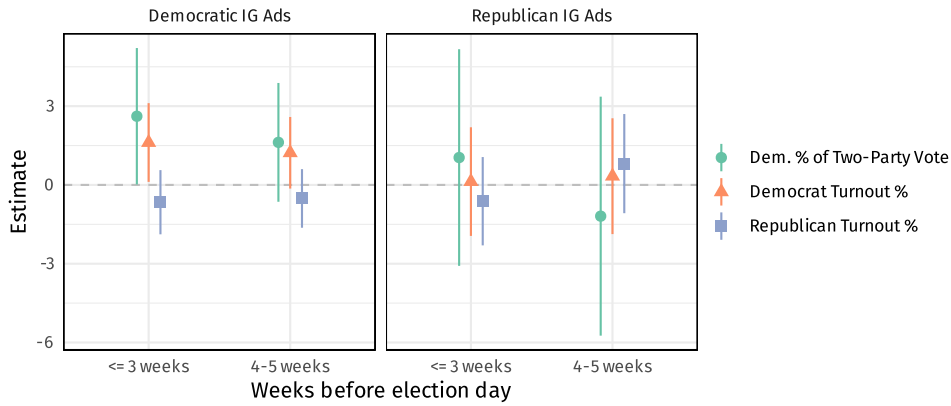

Figure 5. IPTW-FE estimated effects of IG ads by week of the campaign for various outcomes.

6.3 Comparison to other estimated effects

At first glance, the magnitude of the estimated effects may appear large, but recall that we are estimating the effects on down-ballot races for Senate and Governor, where citizens have much less information about the candidates. In 2007, roughly a third of respondents could not name their sitting governor (Hopkins, 2018, p. 67), indicating there may be a considerable number of persuadable voters for these races. Furthermore, the effect sizes we estimate are of similar orders of magnitude to other results in the literature, as we now describe.

Sides et al. (2022), the most comprehensive and credible study of the effects of advertising, uses a difference-in-differences design and a pair-matched border-county design. The latter, in particular, matches counties across market borders and uses matched-pair-year fixed effects to adjust for any election-year unmeasured confounding that affects both counties in a pair similarly. They estimate that an increase of 1000 Democratic ads relative to Republican ads increased the Democratic vote share by 0.38 to 0.87 percentage points depending on the type of race and model specification. Broadly, our analyses reach a similar conclusion to theirs—increases in ads supporting Democratic candidates increases Democratic voteshare—though their estimates are slightly lower than ours. There are key differences between our approaches, however. First, they include all types of ads, not just independent group expenditures. Second, they use as treatment all ads from the last two months together and ignore time-varying confounding within a particular race, treating each race as a point treatment. Ignoring this kind of confounding could lead to bias in the estimated effects if ads were targeted at market areas that had, for example, many negative ads from the opponent (which we adjust for in our weighting model). This kind of confounding could lead to downward bias in their estimates if that negativity hurts the Democratic candidate. Finally, the estimates are difficult to compare directly because their results are in terms of differences in the number of ads for each party, whereas we use the number of weeks with ads as our treatment (and analyse each party separately).

Other studies provide similar effect magnitudes. Hewitt et al. (2024) analysed 146 internal campaign advertising experiments and found that the average advertisement in a down-ballot race had a persuasive effect of 1.2–2.3 percentage points depending on the election year. Analyses of the effect of *Citizens United* provide much larger effect estimates. Klumpp et al. (2016) find the effect of lifting the ban on independent group ads increased the probability of Republicans winning US House elections by 4.1–6.4 percentage points, depending on the specification. Most directly comparable to our study, they found a similar magnitude effect on US Senate elections, though this was statistically insignificant. Their difference-in-differences strategy, however, does not analyse the effect of ads directly, making direct comparisons challenging. Thus, while none of the related literature provides an exact direct comparison of estimated effects, we can say that our magnitudes are within the broad range of values found in the literature.

6.4 Robustness checks

Given that advertising is not randomized across markets, we may worry about residual unmeasured confounding that our approach may miss. To investigate if we can detect any potential biases in our estimation strategy, we use the same designs as above on placebo outcomes. First, we obtain the outcomes for the same media market for the most recent previous election for the same office and use those as outcomes. If our IPTW-FE approach was unable to adjust for unmeasured confounding at the market level, then these estimates would detect bias since future independent group ads cannot affect past electoral outcomes. We also investigate the effects of our estimates on the baseline polling for the Democratic candidate before the five-week period in our MSMs.

Figure 6 shows these results. Both of the IPTW approaches result in estimates very close to zero for all outcomes, which is consistent with our identifying assumptions. Interestingly, the naive approach does show some residual confounding for some of the effects of Democratic group ads. Taken together, these results give us some confidence that further unmeasured confounding is not a major source of bias in our estimates.

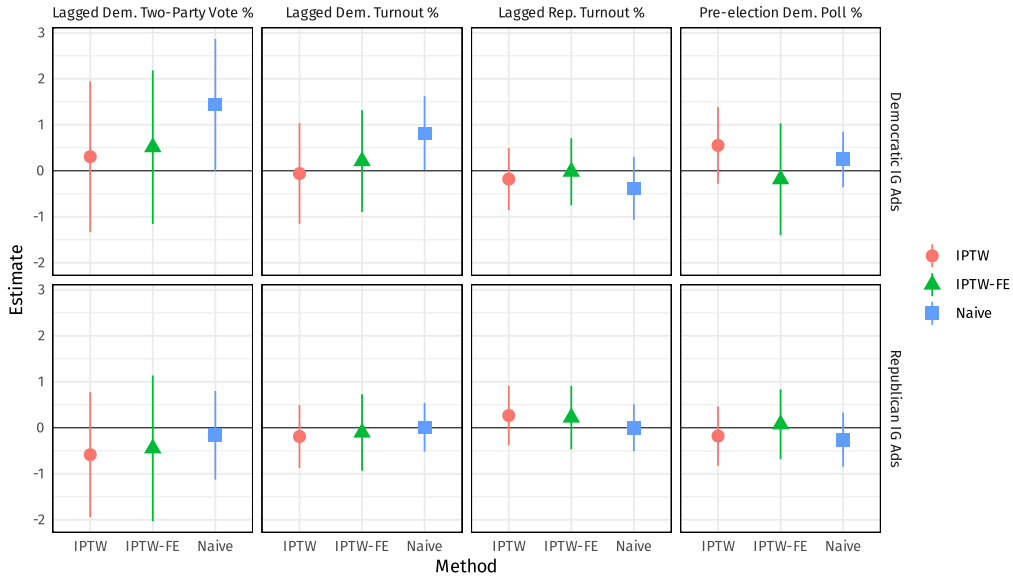

Figure 6. Falsification test results. These are estimated effects on outcomes from the previous election in that market for that office and pre-election polling results.

7 Conclusion

In this paper, we estimated the effects of independent group advertising on electoral outcomes in US state-wide elections. To do so, we developed a method to control for time-constant unmeasured confounding in marginal structural models by using a fixed effects approach to estimate the propensity score of the time-varying treatment. We derived the large-sample properties of this estimator under an asymptotic setup where the number of time periods and the number of units grow together. Simulations showed that the proposed method outperforms a naive approach that omits fixed effects and performs well overall, especially when the magnitude of the heterogeneity is moderate. An obvious place for future research would be to apply these methods to data where we have repeated measurements of the outcomes as well as the treatment. In those situations, it may be possible to develop doubly-robust estimators under fixed effects assumptions.

Acknowledgments

Thanks to Adam Glynn for extensive discussions and feedback. We also thank Dmitry Arkhangelsky, Gary King, and Jacob Montgomery for generous comments. Any errors remain our own.

Conflicts of interest

No competing interests are declared.

Funding

No funding was received for this study.

Author contributions

M.B. and S.Y. contributed equally to this study.

Data availability

The data and code used in this study are available at https://github.com/soichiroy/psfe-replication.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Arellano M., & Hahn J. (2007). Understanding bias in nonlinear panel models: Some recent developments. In R. Blundell, W. Newey, & T. Persson, (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, volume 3 of Econometric Society Monographs* (Chapter 12, pp. 381–409). Cambridge University Press. https://doi.org/10.1017/CBO9780511607547.013

Arkhangelsky D., & Imbens G. W. (2024). Fixed effects and the generalized Mundlak estimator. *Review of Economic Studies*, *91*(5), 2545–2571. https://doi.org/10.1093/restud/rdad089

Baćak V., & Karim M. E. (2019, March). The effect of serious offending on health: A marginal structural model. *Society and Mental Health*, *9*(1), 18–32. https://doi.org/10.1177/2156869318800137

Bacak V., & Kennedy E. H. (2015, February). Marginal structural models: An application to incarceration and marriage during young adulthood. *Journal of Marriage and the Family*, *77*(1), 112–125. https://doi.org/10.1111/jomf.12159

Bang H., & Robins J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–972. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Benoit W. L. (2013, October). *Televised political advertisements*. Oxford University Press. https://doi.org/10.1093/obo/9780199756841-0124

Blackwell M. (2013). A framework for dynamic causal inference in political science. *American Journal of Political Science*, *57*(2), 504–520. https://doi.org/10.1111/j.1540-5907.2012.00626.x

Callaway B., & Sant'Anna P. H. (2021, December). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Cole S. R., & Hernán M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiolology*, *168*(6), 656–664. https://doi.org/10.1093/aje/kwn164

Creamer C. D., & Simmons B. A. (2019). Do self-reporting regimes matter? Evidence from the convention against torture. *International Studies Quarterly: A Publication of the International Studies Association*, *63*(4), 1051–1064. https://doi.org/10.1093/isq/sqz043

Fernández-Val I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, *150*(1), 71–85. https://doi.org/10.1016/j.jeconom.2009.02.007

Fernández-Val I., & Weidner M. (2016). Individual and time effects in nonlinear panel models with large *N*, *T*. *Journal of Econometrics*, *192*(1), 291–312. https://doi.org/10.1016/j.jeconom.2015.12.014

Fernández-Val I., & Weidner M. (2018). Fixed effects estimation of large-*T* panel data models. *Annual Review of Economics*, *10*(1), 109–138. https://doi.org/10.1146/annurev-economics-080217-053542

Fowler E. F., Franz M. M., Ridout T. N., Baum L. M., & Bogucki C. (2019). Political advertising in 2020. https://mediaproject.wesleyan.edu. The Wesleyan Media Project, Department of Government at Wesleyan University.

Gerber A. S., Gimpel J. G., Green D. P., & Shaw D. R. (2011). How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *The American Political Science Review*, *105*(1), 135–150. https://doi.org/10.1017/S000305541000047X

Goldstein K., & Ridout T. N. (2004, May). Measure the effects of televised political advertising in the United States. *Annual Review of Political Science*, *7*(1), 205–226. https://doi.org/10.1146/annurev.polisci.7.012003.104820

Goodman-Bacon A. (2021, December). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Gruber S., Logan R. W., Jarrín I., Monge S., & Hernán M. A. (2015). Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in Medicine*, *34*(1), 106–117. https://doi.org/10.1002/sim.6322

Hahn J., & Kuersteiner G. (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory*, *27*(6), 1152–1191. https://doi.org/10.1017/S0266466611000028

Hahn J., & Newey W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica: Journal of the Econometric Society*, *72*(4), 1295–1319. https://doi.org/10.1111/j.1468-0262.2004.00533.x

Hewitt L., Broockman D., Coppock A., Tappin B. M., Slezak J., Coffman V., Lubin N., & Hamidian M. (2024, November). How experiments help campaigns persuade voters: Evidence from a large archive of campaigns' own experiments. *The American Political Science Review*, *118*(4), 2021–2039. https://doi.org/10.1017/S0003055423001387

Hill S. J., Lo J., Vavreck L., & Zaller J. (2013, October). How quickly we forget: The duration of persuasion effects from mass communication. *Political Communication*, *30*(4), 521–547. https://doi.org/10.1080/10584609.2013.828143

Hopkins D. J. (2018). *The increasingly United States: How and why American political behavior nationalized*. University of Chicago Press.

Huber G. A., & Arceneaux K. (2007). Identifying the persuasive effects of presidential advertising. *American Journal of Political Science*, *51*(4), 957–977. https://doi.org/10.1111/j.1540-5907.2007.00291.x

Imai K., & Kim I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, *63*(2), 467–490. https://doi.org/10.1111/ajps.12417

Imai K., & Ratkovic M. (2015). Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association*, *110*(511), 1013–1023. https://doi.org/10.1080/01621459.2014.956872

Jacobson G. C. (1975, August). The impact of broadcast campaigning on electoral outcomes. *The Journal of Politics*, *37*(3), 769–793. https://doi.org/10.2307/2129324

Kallus N., & Santacatterina M. (2021). Optimal balancing of time-dependent confounders for marginal structural models. *Journal of Causal Inference*, *9*(1), 345–369. https://www.degruyterbrill.com/document/doi/10.1515/jci-2020-0033/html

Kennedy E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, *114*(526), 645–656. https://doi.org/10.1080/01621459.2017.1422737

Klumpp T., Mialon H. M., & Williams M. A. (2016, February). The business of American democracy: *Citizens United* , independent spending, and elections. *The Journal of Law & Economics*, *59*(1), 1–43. https://doi.org/10.1086/685691

Kurer T. (2020). The declining middle: Occupational change, social status, and the populist right. *Comparative Political Studies*, *53*(10–11), 1798–1835. https://doi.org/10.1177/0010414020912283

Ladam C., Harden J. J., & Windett J. H. (2018). Prominent role models: High-profile female politicians and the emergence of women as candidates for public office. *American Journal of Political Science*, *62*(2), 369–381. https://doi.org/10.1111/ajps.12351

MacKinnon J., & White H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325. https://doi.org/10.1016/0304-4076(85)90158-7

Muñoz I. D., & van der Laan M. J. (2011). Super learner based conditional density estimation with application to marginal structural models. *The International Journal of Biostatistics*, *7*(1), 1–20. https://doi.org/10.2202/1557-4679.1356

Neyman J., & Scott E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, *16*(1), 1–32. https://doi.org/10.2307/1914288

Obikane E., Shinozaki T., Takagi D., & Kawakami N. (2018, July). Impact of childhood abuse on suicide-related behavior: Analysis using marginal structural models. *Journal of Affective Disorders*, *234*, 224–230. https://doi.org/10.1016/j.jad.2018.02.034

Ridout T. M., & Franz M. M. (2011). *The persuasive power of campaign advertising*. Temple University Press.

Ridout T. N., Fowler E. F., & Franz M. M. (2021, April). Spending fast and furious: Political advertising in 2020. *The Forum*, *18*(4), 465–492. https://doi.org/10.1515/for-2020-2109

Robins J. M. (1998a). Correction for non-compliance in equivalence trials. *Statistics in Medicine*, *17*(3), 269–302. https://doi.org/10.1002/(ISSN)1097-0258

Robins J. M. (1998b). Marginal structural models. In *1997 Proceedings of the American Statistical Association Section on Bayesian Statistical Science* (pp. 1–10). American Statistical Association.

Robins J. M. (1999). Association, causation, and marginal structural models. *Synthese*, *121*(1/2), 151–179. https://doi.org/10.1023/A:1005285815569

Robins J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran, & D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials, volume 116 of The IMA Volumes in Mathematics and its Applications* (pp. 95–134). Springer-Verlag.

Robins J. M., Hernán M. A., & Brumback B. A. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Rotnitzky A., Lei Q., Sued M., & Robins J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, *99*(2), 439–456. https://doi.org/10.1093/biomet/ass013

Sampson R. J., Laub J. H., & Wimer C. (2006, August). Does marriage reduce crime? A counterfactual approach to within-individual causal effects. *Criminology*, *44*(3), 465–508. https://doi.org/10.1111/j.1745-9125.2006.00055.x

Sharkey P., & Elwert F. (2011, May). The legacy of disadvantage: Multigenerational neighborhood effects on cognitive ability. *The American Journal of Sociology*, *116*(6), 1934–81. https://doi.org/10.1086/660009

Sides J., Vavreck L., & Warshaw C. (2022, May). The effect of television advertising in United States elections. *The American Political Science Review*, *116*(2), 702–718. https://doi.org/10.1017/S000305542100112X

Sobel M. E. (2012). Does marriage boost men's wages?: Identification of treatment effects in fixed effects regression models for panel data. *Journal of the American Statistical Association*, *107*(498), 521–529. https://doi.org/10.1080/01621459.2011.646917

Sun L., & Abraham S. (2021, December). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006

Wodtke G. T., Harding D. J., & Elwert F. (2011, October). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review*, *76*(5), 713–736. https://doi.org/10.1177/0003122411420816

Xiao Y., Moodie E. E., & Abrahamowicz M. (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods*, *2*(1), 1–20. https://doi.org/10.1515/em-2012-0006