Supplementary Appendix for "Estimating Controlled Direct Effects through Marginal Structural Models"

Comparison to Mediation analysis and Structural Nested Mean Models

MSMs overcome limitations that other tools like mediation analysis and structural nested mean models have. For example, causal mediation analysis (Imai, Keele, and Tingley 2010; Pearl 2001) decomposes the total effect of a treatment on an outcome into direct and indirect effects (Imai et al. 2011). However, when one of the confounders is affected by the baseline treatment, mediation analysis is not appropriate because its procedure requires modeling the outcome as a function of treatment history and those problematic confounders affected by the treatment. Therefore, by explicitly conditioning on them we induce post-treatment control bias as explained above (Montgomery, Nyhan, and Torres 2018). More specifically, this method estimates the values of the mediator (the intermediate treatment stage) based on a model that includes relevant confounders and a baseline treatment. Then, the fitted probabilities for each of the values of the treatment are used to predict the outcome. However, for this second step, the model of the outcome includes all treatment stages and all relevant confounders.

Another alternative for the estimation of the ACDE in dynamic settings is the structural nested mean models (SNMMs) approach (Acharya, Blackwell, and Sen 2016; Robins 1997, 1999).[1] SNMMs are a powerful alternative for the estimation of treatment effects especially when the treatments are continuous or comprise a large number of categories (Vansteelandt, Joffe et al. 2014). However, even though SNMMs have the great advantage

[1] For this purpose, these models decompose the overall treatment effect into components that allow for the identification of "demediated" effects.

of working for any type of treatments and confounders, they cannot handle any type of *outcome*. Most SNMMs cannot impose restrictions on the finite support of the outcome (Robins 1999) and are therefore unsuitable for the study of ordinal, multinomial, and count variables.[2] Furthermore, SNMMs are less intuitive and accessible than MSMs and its core concept of "balancing" the sample (Vansteelandt, Joffe et al. 2014). As Acharya, Blackwell, and Sen (2016) indicate, "when the treatment and mediator are binary or only take on a few values, nonparametric or semi-parametric approaches exist to estimating the ACDE, reducing the need for parametric models." In summary, MSMs are accessible, straightforward and often more suitable for the estimation of controlled direct effects when the treatment has few values.

[2] In their paper Acharya, Blackwell, and Sen (2016) present the implementation of SNMMs for continuous variables. Vansteelandt (2010) extends and elaborates on the application of SNMMs to dichotomous outcomes.

Weighting and pseudo-sample

Table SA.1: From Table 1: Calculation of weights for each stratum in sample (full-table)

$Z^{(0)}$	$Z^{(1)}$	$X^{(1)}$	$f(Z^{(1)}\|Z^{(0)})$	$f(Z^{(1)}\|Z^{(0)}, X^{(1)})$	$\mathcal{W}(t)^{-1}$	Original-pop N	Pseudo-pop N
0	0	0	0.600	0.706	0.850	12,000	10,200
0	0	1	0.600	0.462	1.299	6,000	7,794
0	1	0	0.300	0.235	1.277	4,000	5,108
0	1	1	0.300	0.385	0.779	5,000	3,895
0	2	0	0.100	0.059	1.695	1,000	1,695
0	2	1	0.100	0.154	0.649	2,000	1,298
1	0	0	0.367	0.467	0.786	7,000	5,502
1	0	1	0.367	0.267	1.375	4,000	5,500
1	1	0	0.400	0.333	1.201	5,000	6,005
1	1	1	0.400	0.467	0.857	7,000	5,999
1	2	0	0.233	0.200	1.165	3,000	3,495
1	2	1	0.233	0.267	0.873	4,000	3,492
2	0	0	0.200	0.400	0.500	2,000	1,000
2	0	1	0.200	0.100	2.000	1,000	2,000
2	1	0	0.333	0.400	0.832	2,000	1,664
2	1	1	0.333	0.300	1.110	3,000	3,330
2	2	0	0.467	0.200	2.335	1,000	2,335
2	2	1	0.467	0.600	0.778	6,000	4,668

Note: $Z^{(0)}$ is parents' income where 0 is low, 1 is middle and 2 is high. $Z^{(1)}$ is income in adulthood where 0 is low, 1 is middle and 2 is high. $X^{(1)}$ is post-High School education where 0 is no college and 1 is college.

Simulation 1: comparison of bias between weight estimation methods (Section 2.3)

For the comparison of weighting techniques, I simulate a dataset with $n = 1,000$ where the main outcome of interest is attendance of a rally (0=No, 1=Yes). The data includes two relevant sequences of covariates: parents' income and income in adulthood (the treatment sequence), and whether parents and respondent attended college (the confounders sequence). In this setup, college attendance of a subject acts as a confounder of income (second stage of the treatment) and rally attendance, but is also affected by parents' income (baseline treatment). The parameters are tuned to purposely allow for the possibility of observing samples in which the positivity assumption is not fulfilled. This is, there are combinations of the sequence treatment and college attendance that do not have any observations. The parameters and specification of the simulation are presented below. The idea behind this specification is to illustrate the advantages of MSMs over traditional models even when one of the main assumptions that the former requires are mildly violated. Second, I use this data to estimate and record the ACDEs from four different models: the saturated or *naïve* model, and three MSMs that use weights calculated using ologit, GAM and RF models.

The specification of the outcome model is the following:

$$Pr(Y = 1|\overrightarrow{\boldsymbol{Z}}) = \text{logit}^{-1}(\alpha_0 + \boldsymbol{Z^{(0)}\beta} + \boldsymbol{Z^{(1)}\gamma} + \boldsymbol{(Z^{(0)} \times Z^{(1)})\delta}) \tag{1}$$

The rest of the parameters are specified in the following way:

Baseline covariate: parents' college attendance

$$X_i^{(0)} \sim \mathcal{B}ernoulli(1, 0.5)$$

Baseline treatment: parents' income

$$
\begin{aligned}
Z_i^{(0)} &\sim \mathcal{C}ategorical(3, \mathbf{p}_i) \\
\mathbf{p}_i &= (F(\eta_{i1}), F(\eta_{i2}) - F(\eta_{i1}), 1 - F(\eta_{i2})) \\
F(\eta_{ik}) &= \frac{\exp(\theta_k - \eta_{ik})}{1 + \exp(\theta_k - \eta_{ik})} \\
\eta_{ik} &= -2.5 + 1.5X_i^{(0)} + u_i \\
u_i &\sim \mathcal{N}(0, 0.5) \\
(\theta_1, \theta_2) &= (-1.25, 0.45)
\end{aligned}
$$

Covariate affected by baseline treatment: subject's college attendance

$$
\begin{aligned}
X_i^{(1)} &\sim \mathcal{B}ernoulli(1, p_i^\dagger) \\
p_i^\dagger &= \frac{\exp(-2.5 + 0.5X_i^{(0)} + 1.1Z_i^{(0)} + u_i^\dagger)}{1 + \exp(-2.5 + 0.5X^{(0)} + 1.1Z^{(0)} + u_i^\dagger)} \\
u_i^\dagger &\sim \mathcal{N}(0, 0.35)
\end{aligned}
$$

Second-stage treatment: subject's income

$$
\begin{aligned}
Z_i^{(1)} &\sim \mathcal{C}ategorical(3, \mathbf{p}_i^*) \\
\mathbf{p}_i^* &= (F(\eta_{i1}^*), F(\eta_{i2}^*) - F(\eta_{i1}^*), 1 - F(\eta_{i2}^*)) \\
F(\eta_{ik}^*) &= \frac{\exp(\theta_k^* - \eta_{ik}^*)}{1 + \exp(\theta_k^* - \eta_{ik}^*)} \\
\eta_{ik}^* &= -3.5 + 0.2X_i^{(0)} + 1Z_i^{(0)} + 0.6X_i^{(1)} + u_i^* \\
u_i^* &\sim \mathcal{N}(0, 0.5) \\
(\theta_1^*, \theta_2^*) &= (-1.05, 0.65)
\end{aligned}
$$

Outcome: participation in a rally

$$
\begin{aligned}
&\mathbf{Y}_i \sim \mathcal{B}ernoulli(p_i^+) \\
&p_i^+ = \frac{\exp(\eta_i^+)}{1+\exp(\eta_i^+)} \\
&\eta_i^+ = -3 + 0.2X_i^{(0)} + 1.5Z_i^{(0)} + 0.4X_i^{(1)} + 0.2Z^{(1)} + u_i^+ \\
&u_i^+ \sim \mathcal{N}(0, 0.4)
\end{aligned}
$$

For this exercise, I calculated nine potential outcomes according to the multiple combinations of the baseline and second-stage treatment values. The *true* controlled direct effects of the baseline treatment are calculated for each individual using this framework. The results are presented in Figure 3 in the manuscript.

Simulation 2

The following simulations illustrate the advantages that marginal structural models have over saturated models that control for post-treatment confounders under several conditions.

I conduct three sets of simulations, each changing the value of one of the following parameters while keeping the others constant: number of observations (n), the effect of a covariate on the treatment sequence, $X^{(0)}$, and the effect of a confounder of treatment and outcome affected by the baseline treatment. In each set I generate a set of variables with the structure presented below: a baseline treatment stage with three values $Z^{(0)}$, a binary covariate $X^{(1)}$ affected by the baseline treatment, an intermediate treatment stage affected by both $X^{(0)}$ and $Z^{(0)}$, an outcome Y generated by all of these variables. The third simulation also includes another binary covariate $W^{(1)}$ affected by the baseline treatment. Covariates $X^{(0)}$, $W^{(1)}$ and $X^{(1)}$ confound the relationship between the outcome and the treatment stages.

Baseline covariate

$$X_i^{(0)} \sim \mathcal{Bernoulli}(1, 0.4)$$

Baseline treatment

$$\begin{aligned}
Z_i^{(0)} &\sim \mathcal{Categorical}(3, \mathbf{p}_i) \\
\mathbf{p}_i &= (F(\eta_{i1}), F(\eta_{i2}) - F(\eta_{i1}), 1 - F(\eta_{i2})) \\
F(\eta_{ik}) &= \frac{\exp(\theta_k - \eta_{ik})}{1 + \exp(\theta_k - \eta_{ik})} \\
\eta_{ik} &= -2.5 + \beta_1 X_i^{(0)} + u_i \\
u_i &\sim \mathcal{N}(0, 0.5) \\
(\theta_1, \theta_2) &= (-1.25, 0.05)
\end{aligned}$$

When held constant $\beta_1 = 1$.

Covariates affected by baseline treatment

$$\begin{aligned}
X_i^{(1)} &\sim \mathcal{B}ernoulli(1, p_i^\dagger) \\
p_i^\dagger &= \frac{\exp(\eta_i^\dagger)}{1 + \exp(\eta_i^\dagger)} \\
\eta_i^\dagger &= -2.5 + 0.5X_i^{(0)} + 1.1Z_i^{(0)} + u_i^\dagger \\
u_i^\dagger &\sim \mathcal{N}(0, 0.035)
\end{aligned}$$

$$\begin{aligned}
W_i^{(1)} &\sim \mathcal{B}ernoulli(1, p_i^\ddagger) \\
p_i^\ddagger &= \frac{\exp(\eta_i^\ddagger)}{1 + \exp(\eta_i^\ddagger)} \\
\eta_i^\ddagger &= -2.5 + 0.5X_i^{(0)} - 1.5Z_i^{(0)} + u_i^\ddagger \\
u_i^\ddagger &\sim \mathcal{N}(0, 0.035)
\end{aligned}$$

Second-stage treatment

$$\begin{aligned}
Z_i^{(1)} &\sim \mathcal{C}ategorical(3, \mathbf{p}_i^*) \\
\mathbf{p}_i^* &= (F(\eta_{i1}^*), F(\eta_{i2}^*) - F(\eta_{i1}^*), 1 - F(\eta_{i2}^*)) \\
F(\eta_{ik}^*) &= \frac{\exp(\theta_k^* - \eta_{ik}^*)}{1 + \exp(\theta_k^* - \eta_{ik}^*)} \\
\eta_{ik}^* &= -3.5 + \beta_2 X_i^{(0)} + 1Z_i^{(0)} + 0.6X_i^{(1)} + \gamma_1 W_i^{(1)} + u_i^* \\
u_i^* &\sim \mathcal{N}(0, 0.5) \\
(\theta_1^*, \theta_2^*) &= (-1.05, 0.65)
\end{aligned}$$

When held constant, $\beta_2 = 0.5$ and $\gamma_1 = 0.6$.

Outcome

$$
\begin{aligned}
&\mathbf{Y}_i \sim \mathcal{B}ernoulli(p_i^+) \\
&p_i^+ = \frac{\exp(\eta_i^+)}{1+\exp(\eta_i^+)} \\
&\eta_i^+ = -3 + 0.2X_i^{(0)} + \gamma_2 W_i^{(1)} + 1.5Z_i^{(0)} + 0.4X_i^{(1)} + 0.2Z^{(1)} + u_i^+ \\
&u_i^+ \sim \mathcal{N}(0, 0.4)
\end{aligned}
$$

When held constant, $\gamma_2 = 0.3$.

The three sets of simulations have the main objective of comparing and analyzing the biases in the estimation of controlled direct effects in situations where the sequential ignorability and positivity assumptions required by MSMs are violated.

The first simulation varies the number of observations in the simulated datasets (from 60 to 2,000). The objective of this exercise is to explore the sample size properties of the IPTW estimator while also setting scenarios, such as those with very few observations, where the positivity assumption is likely to be violated. When held constant in the rest of the simulations, $n = 1,000$.

The second simulation varies the effect of a confounder $X^(0)$ on the treatment assignment (both the effect of $X^{(0)}$ on $Z^{(0)}$ [denoted by β_1] and $Z^{(1)}$ [denoted by β_2]). This exercise also helps to illustrate the bias that arises in cases where the treatment assignment is heavily unbalanced and therefore causing 1) certain covariate and treatment histories to be empty and/or 2) to obtain extreme weights for some combinations of such variables.

Finally, the third simulation explores the magnitude and variance of the bias when the researcher omits a confounder of the second stage of the treatment and the outcome. In the simulation, I increase the importance of such confounder by varying the impact of $W^{(1)}$ on the treatment $Z^{(1)}$ (denoted by γ_1) and on the outcome Y (denoted by γ_2).

In order to assess and compute the bias for each case, I generate a set of *potential* *outcomes* to calculate the "true" controlled direct effects of $Z^{(0)}$ on Y. There are nine CDEs which I present in Table SA.2. Then, for each set of simulations I estimated the CDEs based on the *observed outcomes* using two modeling strategies: a marginal structural model (MSM) which implies a weighted regression of the outcome on the two treatments using weights estimated through IPTW, and a saturated model which includes confounders affected by the treatment. For illustrative purposes, weights were estimated using categorical logistic regressions.

Table SA.2: Simulated controlled direct effects

CDE	$Y_{Z^{(0)}=a,Z^{(1)}=b} - Y_{Z^{(0)}=a',Z^{(1)}=b}$
1	$Y_{10} - Y_{00}$
2	$Y_{20} - Y_{00}$
3	$Y_{20} - Y_{10}$
4	$Y_{11} - Y_{01}$
5	$Y_{21} - Y_{01}$
6	$Y_{21} - Y_{11}$
7	$Y_{12} - Y_{02}$
8	$Y_{22} - Y_{02}$
9	$Y_{22} - Y_{12}$

For each of the varying values of the parameters of interest, I simulate 400 datasets. After collecting the relevant estimates of ACDEs from a MSM and a saturated model in each dataset, I take the difference between such estimates and the true controlled direct effects. These values represent a measure of bias. Figure SA.1 below shows three panels with the distributions of bias when estimating CDE number 1 using either a MSM or a saturated model under different conditions. In each panel, the y-axis indicates the magnitude of the bias in the estimation of CDE 1, while the upper and lower x-axes show the different values that the parameter of interests take in each set of simulated datasets. The lines indicate the mean bias: red and dashed for the saturated model, and bold and black for the MSM. The gray areas indicate the 5^{th} and 95^{th} percentiles in the distribution of bias.

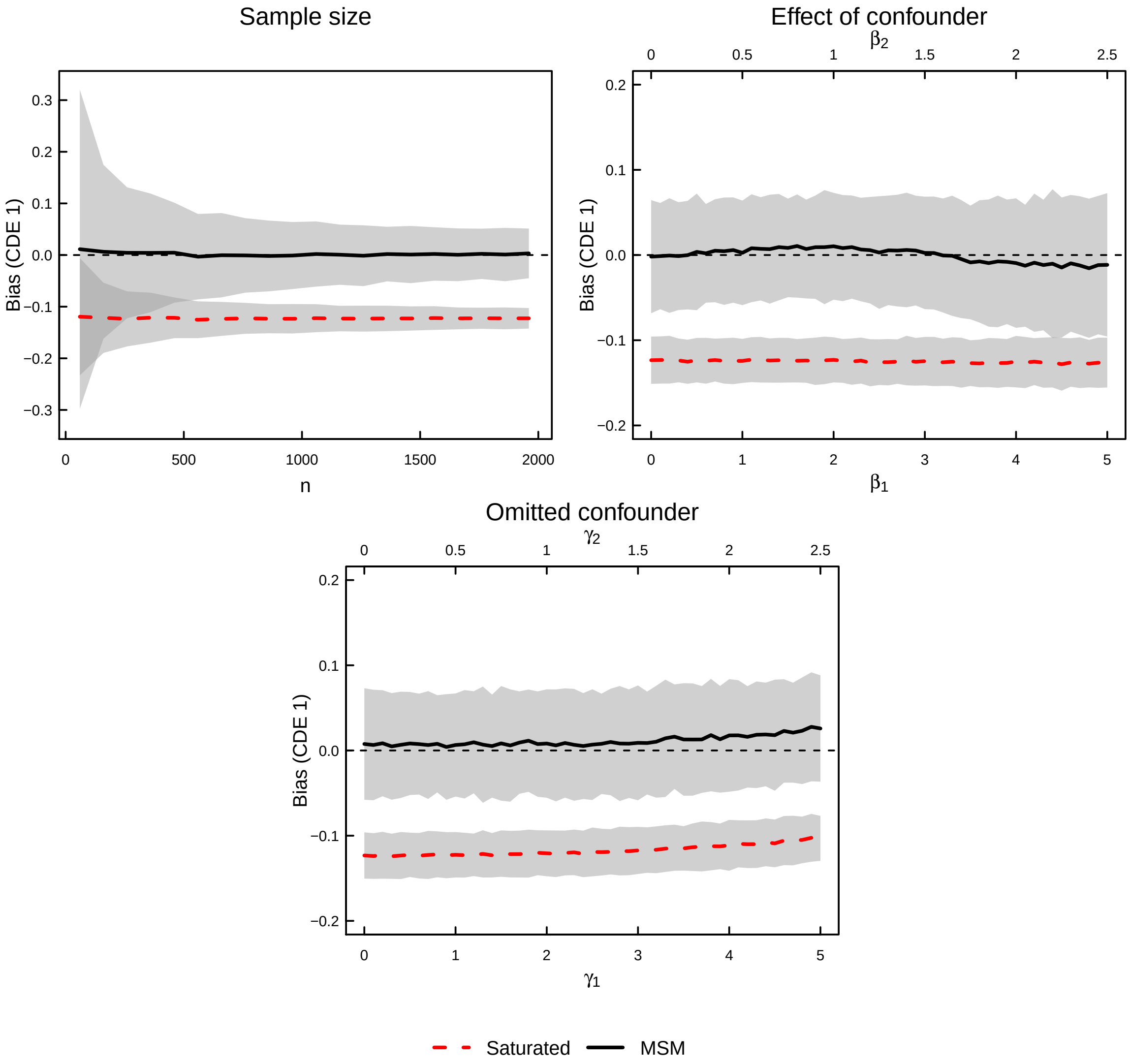
The results for the first simulation varying the sample size show that although MSMs slightly overestimate the real CDE 1 in small samples (with around 60 to 100 observations), the bias quickly converges to 0 and stays steady for large sample sizes. However, the estimate of ACDE 1 using the saturated model remains biased even when sample size increases. Controlling for the covariate affected by the baseline treatment originates this bias. Although the average bias is close to 0 when implementing a MSM, it is important to consider that its variance is 1) higher than the variance of bias from the saturated model, and 2) decreases at the same time as sample size increases. In general, saturated models perform better in terms of standard errors. A result that does not come as a surprise given the weighting process involved in the estimation of ACDEs in a MSM framework.

The second simulation increases the importance of one of the pre-treatment covariates affecting the assignment of both stages of treatment. Increasing the effect of $X^{(0)}$ on $Z^{(0)}$ and $Z^{(1)}$ yields the following results. Under this setting, we find that, as expected, the estimates of ACDE 1 using MSMs start unbiased when the effect of the covariate is 0 or about 0.5 (which once plugged in a the link function it represents a substantive effect in terms of probabilities). When the effects increase, we observe that the average bias departs from zero in both negative and positive directions. This bias, however, is still smaller than the almost constant and negative bias in the estimates that a saturated model yield. Further, we observe an increase in the variance of the bias distribution of MSMs as the effect of $X^{(0)}$ becomes more important. And even in the case where the effect is zero, the variance of the MSM bias is significantly larger than the one for the saturated model.

Finally, the third simulation shows the distributions of bias when a covariate affected by the baseline treatment gets stronger AND when it is omitted from both the MSM and saturated models. In this case, we observe that the estimator is nearly unbiased when this effect is zero, but has a positive trend departing from zero. However, this bias is consistently lower than the one yielded by the saturated model who shows a decreasing trend in terms of bias. While it might appear counterintuitive, this trend can be explained by the "accumulation" of different biases and the bias trade-off that was explained in Section 1 of this text: while ignoring an increasingly strong confounder can have pernicious consequences as the bold line in this simulation shows, including it may also be problematic. In this cases, it seems that the bias generated by post-treatment control is higher than the confounding bias, and therefore we observe a still biased but improving trend in the results.

The main conclusions that we derive from this exercise is that 1) MSMs perform significantly better than saturated models in terms of bias, but 2) the variance of the bias of MSMs suggests a less efficient estimator. This is consistent with the results found by Westreich et al. 2012 in which they also conduct a set of simulations to compare bias, standard errors and mean squared errors. The evaluation of whether the increased in variance that comes from weighting is outweighed by the reduction in bias that MSMs offer heavily depends on the particular characteristics of the study: the distribution of treatments, the effect of the confounders, etc. For example, if the post-treatment confounder has a very small effect on the treatment sequence and outcome, then a small bias is preferred to large variances. However, the simulations above, conducted under different settings, suggest that the increased variance associated with weighting versus over-adjusting is not too costly. As the graphs show, the MSMs provide a much better coverage of the real estimates than saturated models even at the tail of the bias distribution.

Figure SA.1: Distribution of bias: MSMs vs saturated models



Note: Based on 400 simulated datasets per value.

Application

Data description

The sample framework of the Youth-Parent Socialization Panel is composed of senior High School students in 1965. For this wave, the data comes from a nationally representative sample of 1,669 students distributed across 97 public and nonpublic schools selected with probability proportional to size. In the 1965 wave the parents of the students were also interviewed. For the majority of the students, either one or both parents were interviewed. However, for a small number of cases, no parent was interviewed. For the 1973, 1982 and 1997 waves, students were recontacted and resurveyed. Although most of the surveys were completed face-to-face, a number of them in the follow-up waves were completed through mail interviews and computer-assisted telephone interviews (CATI).

For the first treatment stage, these covariates include education of both mother and father, and race and level of interest in politics of the head of the household. These measures were collected from the parents in 1965. Only in those cases were there was no information available either from the mother or the father, I use the student's answers to those questions.

For the second stage, the confounders are the student's characteristics such as education, political interest, political efficacy and political knowledge as indicators of political skills, motivations and self-confidence. For the full model, I include gender and race of the student as "non-problematic" confounders given that income cannot affect these variables.

Wording

Outcome variables

- Attend a rally
  - Question: Have you gone to any political meetings, rallies, dinners, or other things like that since 1973?
  - Answers: Yes, No
- Donate money

  - Question: Have you given any money or bought any stickers to help a particular party, candidate, or group pay campaign expenses since 1973?
  - Answer: Yes, No

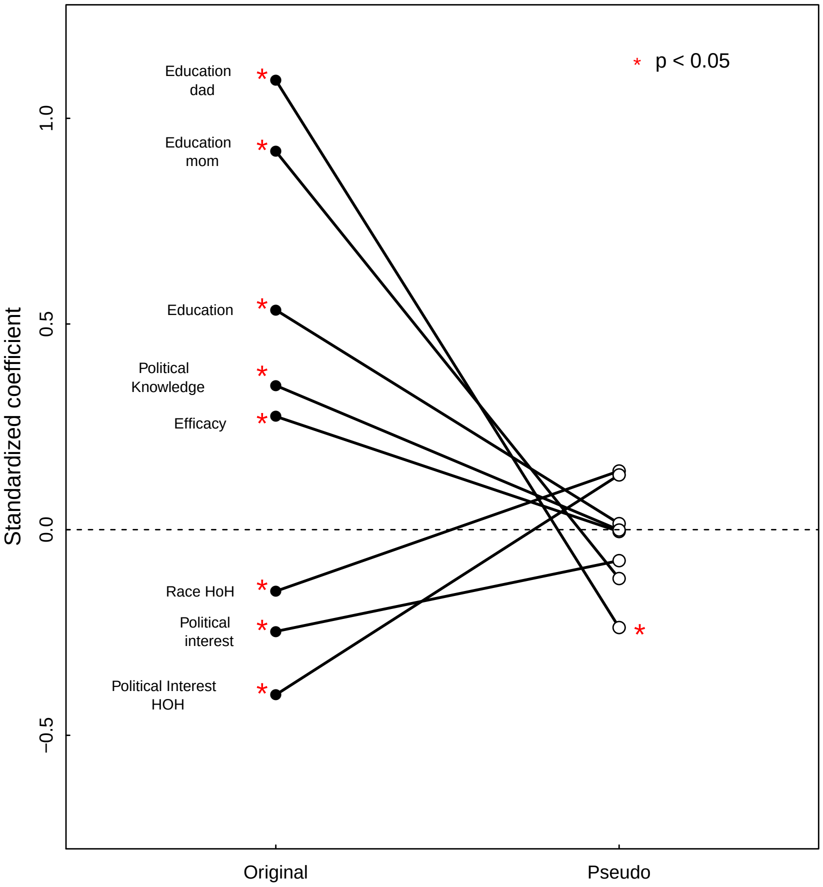
Treatment variables

- Parents' income: Quartiles based on the categories answered by student's parents in 1965
  - Question: About what do you think your total income will be this year for yourself and your immediate family?
- Income in adulthood: Quartiles based on the categories answered by student in 1982
  - Question: Please look at this page and tell me the letter of the income group that includes the income of all members of your family living here in 1981 before taxes. This figure should include salaries, wages, pensions, dividends, interest, and all other income. If uncertain: what would be your best guess?

Extended analysis

The weights estimated from Equation 12 in the main text aim to balance the second stage of the treatment, income in adulthood, across confounders. Figure SA.2 shows that the weights lead to a more balanced sample. This figure illustrates the difference in the standardized coefficients of the confounders on income in adulthood in the original population (left side) and the pseudo-population (right side). The figure shows that while in the original population all covariates significantly predict levels of income in adulthood, in the pseudo-population, almost all of these are no longer significantly associated with the latter. In other words, we successfully "broke" the link between post-treatment confounders and treatment.

Figure SA.2: Balancing covariates



References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3): 512–29.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological methods* 15(4): 309–334.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." *American Political Science Review* 105(4): 765–789.

Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3): 760–775.

Pearl, Judea. 2001. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc. pp. 411–420.

Robins, James M. 1997. "Causal inference from complex longitudinal data." In *Latent variable modeling and applications to causality*. Springer pp. 69–117.

Robins, James M. 1999. "Marginal Structural Models versus Structural Nested Models as tools for causal inference." In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Vol. 116 New York: Springer-Verlag pp. 95–133.

Vansteelandt, Stijn. 2010. "Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models." *Biometrika* pp. 1–14.

Vansteelandt, Stijn, Marshall Joffe et al. 2014. "Structural nested models and G-estimation: The partially realized promise." *Statistical Science* 29(4): 707–731.