# Supplemental Online Materials for 'Enhancing causal pursuits in organizational science: Targeting the effect of treatment on the treated in research on vulnerable populations'

Wen Wei Loh, and Dongning Ren

March 18, 2024

## A Non-parametric identification of the ETT

When both assumptions (5) and (12) are met, we can consistently estimate the ETT using the observable data as follows:

$$\begin{aligned}
\text{ETT} &\equiv \text{E}(Y^1 - Y^0 | X = 1) \\
&= \sum_c \left\{ \text{E}(Y^1 | X = 1, C = c) - \text{E}(Y^0 | X = 1, C = c) \right\} \text{Prob}(C = c | X = 1) \\
&= \sum_c \left\{ \text{E}(Y | X = 1, C = c) - \text{E}(Y^0 | X = 1, C = c) \right\} \text{Prob}(C = c | X = 1) \\
&= \sum_c \left\{ \text{E}(Y | X = 1, C = c) - \text{E}(Y^0 | X = 0, C = c) \right\} \text{Prob}(C = c | X = 1) \\
&= \sum_c \left\{ \text{E}(Y | X = 1, C = c) - \text{E}(Y | X = 0, C = c) \right\} \text{Prob}(C = c | X = 1).
\end{aligned}$$

The first equality follows from the linearity of the expectation operator and the law of iterated expectations; the second and fourth equalities follow from (12); and the third equality follows from weak ignorability in (5). The ETT can thus be interpreted as the weighted average difference in outcomes between those who experienced discrimination and those who did not, within each stratum $C = c$, i.e., $\text{E}(Y | X = 1, C = c) - \text{E}(Y | X = 0, C = c)$, with weights being the proportions of each stratum among those who experienced discrimination, i.e., $\text{Prob}(C = c | X = 1)$.

# B Why the ETT cannot generally be estimated using an estimator of the ATE

When both assumptions (5) and (12) are met, we showed in (6) of the main text that the ETT can be consistently estimated using conditional expectations of the observable quantities:

$$\text{ETT} = \sum_c \{\text{E}(Y|X=1, C=c) - \text{E}(Y|X=0, C=c)\} \, \text{Prob}(C=c|X=1).$$

When the stronger assumption (4) is further satisfied, the ATE can be consistently estimated as:

$$\begin{aligned}
\text{ATE} &\equiv \text{E}(Y^1 - Y^0) \\
&= \sum_c \big\{\text{E}(Y^1|C=c) - \text{E}(Y^0|C=c)\big\} \, \text{Prob}(C=c) \\
&= \sum_c \big\{\text{E}(Y^1|X=1, C=c) - \text{E}(Y^0|X=0, C=c)\big\} \, \text{Prob}(C=c) \\
&= \sum_c \{\text{E}(Y|X=1, C=c) - \text{E}(Y|X=0, C=c)\} \, \text{Prob}(C=c).
\end{aligned}$$

The first equality follows from the linearity of the expectation operator and the law of iterated expectations; the second equality follows from the strong ignorability assumption (4); and the third equality follows from counterfactual consistency (12).

We now use a simple scenario to illustrate why the ETT cannot generally be estimated using an estimator of the ATE. Let $C$ be a binary variable that takes a value of either 0 or 1. As an example, following the first illustration in the main text, let $C$ indicate whether an individual is employed in the information technology sector ($C = 1$) or not ($C = 0$).[1] Suppose that the conditional mean outcome given $X$ and $C$ is correctly specified as:

$$\text{E}(Y|X, C) = \beta_0 + \beta_x X + \beta_{xc} XC + \beta_c C.$$

Plugging this expression into (6) yields an estimator of the ETT as $\beta_x + \beta_{xc} \, \text{Prob}(C = 1|X = 1) = \beta_x + \beta_{xc} \, \text{E}(C|X = 1)$. In contrast, the estimator of the ATE is $\beta_x + \beta_{xc} \, \text{Prob}(C = 1) = \beta_x + \beta_{xc} \, \text{E}(C)$. Therefore, the estimators will generally differ, except when $X$ and $C$ are independent so that $\text{E}(C|X = 1) = \text{E}(C)$, or there are no interactions between $X$ and $C$ so that $\beta_{xc} = 0$, as stated in the main text.

It is also worth noting that merely comparing the mean observed outcomes between the $X = 1$ and $X = 0$ groups does not yield a consistent estimator of the ETT. This is

[1] We thank anonymous reviewer 1 for encouraging us to include an example for binary $C$.

because:

$$
\begin{aligned}
&\mathrm{E}(Y|X=1) - \mathrm{E}(Y|X=0) \\
&= \sum_{c=0}^{1} \mathrm{E}(Y|X=1, C=c)\,\mathrm{Prob}(C=c|X=1) - \mathrm{E}(Y|X=0, C=c)\,\mathrm{Prob}(C=c|X=0) \\
&= \sum_{c=0}^{1} \{\beta_0 + \beta_x + (\beta_{xc} + \beta_c)c\}\,\mathrm{Prob}(C=c|X=1) - (\beta_0 + \beta_c c)\,\mathrm{Prob}(C=c|X=0) \\
&= \sum_{c=0}^{1} (\beta_x + \beta_{xc}c)\,\mathrm{Prob}(C=c|X=1) + (\beta_0 + \beta_c c)\,\{\mathrm{Prob}(C=c|X=1) - \mathrm{Prob}(C=c|X=0)\} \\
&= \{\beta_x + \beta_{xc}\,\mathrm{Prob}(C=1|X=1)\} + \sum_{c=0}^{1} (\beta_0 + \beta_c c)\,\{\mathrm{Prob}(C=c|X=1) - \mathrm{Prob}(C=c|X=0)\}\,.
\end{aligned}
$$

When $X$ is non-randomized so that $\mathrm{Prob}(C=c|X=1) \neq \mathrm{Prob}(C=c|X=0)$, then the sum in the last term is non-zero and it follows that the difference in mean outcomes $\mathrm{E}(Y|X=1) - \mathrm{E}(Y|X=0)$ will not be consistent for the ETT.

# C Estimators of the ETT using observable quantities

In this section, we derive the closed-form expressions of the different estimators of the ETT used in the main text. The derivations follow those in Moodie et al. (2018). When both assumptions (12) and (5) are met, the ETT causal estimand can be expressed in terms of the observable data as:

$$
\begin{aligned}
\mathrm{ETT} &= \sum_{c} \{\mathrm{E}(Y|X=1, C=c) - \mathrm{E}(Y|X=0, C=c)\}\,\mathrm{Prob}(C=c|X=1) \\
&= \mathrm{E}(Y|X=1) - \sum_{c} \mathrm{E}(Y|X=0, C=c)\,\mathrm{Prob}(C=c|X=1).
\end{aligned}
$$

The first term of the ETT is the conditional mean outcome among those who experienced discrimination ($X=1$) and can be readily estimated using the sample analog, i.e.,

$$
\hat{\mathrm{E}}(Y|X=1) = \left(\sum_i X_i\right)^{-1} \sum_i X_i Y_i, \tag{S1}
$$

where an $i$ subscript indexes each individual.

Different estimators are obtained by utilizing different approaches to estimate the second term of the ETT. Suppose we use a regression function to model the conditional mean outcome in terms of the covariates $C$, among those who did not experience discrimination ($X=0$). Denote the estimated regression model by $\mu^0(C) = \hat{\mathrm{E}}(Y|X=0, C)$, where the

notation emphasizes that this is a function only of $C$ and is conditional on $X = 0$. An example is provided in (9) of the main text. An estimator of the second term of the ETT is thus:

$$\sum_c \mu^0(c)\widehat{\text{Prob}}(C = c|X = 1) = \sum_i \mu^0(C_i)\frac{X_i}{\sum_i X_i} = \left(\sum_i X_i\right)^{-1} \sum_i X_i \mu^0(C_i). \quad \text{(S2)}$$

An estimator of the ETT which combines the estimator of the first term in (S1) and the estimator of the second term in (S2) is thus:

$$\widehat{\text{ETT}}_{\text{Y}} = \left(\sum_i X_i\right)^{-1} \sum_i X_i \left\{Y_i - \mu^0(C_i)\right\}.$$

This is the outcome-only regression estimator stated in (S9) in section E. It relies on correctly specifying the outcome model $\mu^0(C)$.

We now derive a different estimator that uses a propensity score model instead of an outcome model. We can write the second term of the ETT as:

$$\begin{aligned}
&\sum_c \text{E}(Y|X = 0, C = c)\,\text{Prob}(C = c|X = 1) \\
&= \sum_c \text{E}(Y|X = 0, C = c)\frac{\text{Prob}(C = c|X = 1)}{\text{Prob}(C = c|X = 0)}\,\text{Prob}(C = c|X = 0) \\
&= \sum_c \text{E}(Y|X = 0, C = c)\,\text{Prob}(C = c|X = 0)\frac{\text{Prob}(X = 1|C = c)\,\text{Prob}(X = 0)}{\text{Prob}(X = 0|C = c)\,\text{Prob}(X = 1)} \\
&= \frac{1}{\text{Prob}(X = 1)} \sum_c \text{E}(Y|X = 0, C = c)\,\text{Prob}(C = c, X = 0)\frac{\text{Prob}(X = 1|C = c)}{\text{Prob}(X = 0|C = c)} \\
&= \frac{1}{\text{Prob}(X = 1)} \sum_c \sum_y y\,\text{Prob}(Y = y, C = c, X = 0)\frac{\text{Prob}(X = 1|C = c)}{1 - \text{Prob}(X = 1|C = c)}.
\end{aligned}$$

Let $W^0(C) = \dfrac{\text{Prob}(X = 1|C)}{1 - \text{Prob}(X = 1|C)}$ denote the conditional odds of experiencing discrimination given the covariates $C$. Note that this motivates the weights $W$ among those who did not experience discrimination ($X = 0$) as defined in the main text. An estimator of the second term of the ETT is thus:

$$\frac{1}{\text{Prob}(X = 1)} \sum_c \sum_y y\widehat{\text{Prob}}(Y = y, C = c, X = 0)W^0(c) = \left(\sum_i X_i\right)^{-1} \sum_i Y_i(1 - X_i)W_i. \quad \text{(S3)}$$

Recall that the weight $W$ equaled 1 for those who experienced discrimination ($X = 1$). Hence, we can rewrite an estimator of the first term of the ETT as:

$$\hat{\mathrm{E}}(Y|X=1) = \left(\sum_i X_i\right)^{-1} \sum_i X_i W_i Y_i. \tag{S4}$$

An estimator of the ETT which combines the estimator of the first term in (S4) and the estimator of the second term in (S3) is thus:

$$\widehat{\mathrm{ETT}}_{\mathrm{IPW}} = \left(\sum_i X_i\right)^{-1} \sum_i \{X_i - (1 - X_i)\} W_i Y_i.$$

This is the IPW estimator stated in (S10) in section E. It relies on correctly specifying the propensity score model $\mathrm{Prob}(X = 1|C)$ and, subsequently, the weights $W$ among those with $X = 0$.

Finally, we combine both estimators to construct the doubly robust estimator introduced in (10) of the main text. We can write the second term of the ETT as follows:

$$\begin{aligned}\mathrm{E}(Y^0|X=1) &= \mathrm{E}\left\{Y^0 - \mu^0(C) + \mu^0(C)|X=1\right\} \\ &= \mathrm{E}\left\{Y^0 - \mu^0(C)|X=1\right\} + \mathrm{E}\left\{\mu^0(C)|X=1\right\}.\end{aligned}$$

An estimator of the second term of the ETT can be obtained by replacing $Y_i$ in (S3) and (S4) with $Y_i - \mu^0(C_i)$ and $\mu^0(C_i)$, respectively; i.e.,

$$\left(\sum_i X_i\right)^{-1} \sum_i \{Y_i - \mu^0(C_i)\}(1 - X_i) W_i + \left(\sum_i X_i\right)^{-1} \sum_i X_i W_i \mu^0(C_i). \tag{S5}$$

An estimator of the ETT which combines the estimator of the first term in (S4) and the estimator of the second term in (S5) is thus:

$$\begin{aligned}\widehat{\mathrm{ETT}} &= \left(\sum_i X_i\right)^{-1} \sum_i X_i W_i Y_i \\ &\quad - \left(\sum_i X_i\right)^{-1} \sum_i \{Y_i - \mu^0(C_i)\}(1 - X_i) W_i - \left(\sum_i X_i\right)^{-1} \sum_i X_i W_i \mu^0(C_i) \\ &= \left(\sum_i X_i\right)^{-1} \sum_i X_i W_i \left\{Y_i - \mu^0(C_i)\right\} - (1 - X_i) W_i \left\{Y_i - \mu^0(C_i)\right\} \\ &= \left(\sum_i X_i\right)^{-1} \sum_i \{X_i - (1 - X_i)\} W_i \left\{Y_i - \mu^0(C_i)\right\} \\ &= \left(\sum_i X_i\right)^{-1} \sum_i \{X_i - (1 - X_i)\} W_i D_i.\end{aligned}$$

This is the estimator in (10) of the main text. It is a so-called augmented inverse probability of treatment weighted (AIPW) estimator that augments the IPW estimator with an outcome model; specifically, it replaces $Y_i$ in $\widehat{\text{ETT}}_{\text{IPW}}$ with $D_i = Y_i - \mu^0(C_i)$.

The estimator is doubly robust in the sense that it will be consistent for the ETT when either the propensity score model or the outcome model is correctly specified, even if the other is incorrectly specified, without having to know which is (in)correct (Moodie et al., 2018). Readers are referred to Glynn and Quinn (2010), Robins et al. (2007), and Schafer and Kang (2008) for explanations of the statistical theory underpinning doubly robust AIPW estimators when targeting the ATE.

# D Conditional ETTs for specific subgroups

In practice, research interest is often in investigating the heterogeneity of the impact of discrimination across different subgroups. Here, we show that the ETT causal estimand can be readily used to unveil heterogeneous effects. The ETT can vary across subgroups or strata defined by a pre-specified subset of the baseline covariates $C$, henceforth denoted by $Z \subseteq C$. Suppose interest is in the conditional effect of discrimination among those with a particular value of $Z = z$. Then, the conditional ETT (CETT) given $Z = z$ is readily defined as:

$$\text{CETT}(z) \equiv \text{E}(Y^1 - Y^0 | X = 1, Z = z). \tag{S6}$$

In other words, the CETT($z$) is the average difference in potential outcomes among those who perceived discrimination ($X = 1$) and with values of $Z = z$. Zang et al. (2023) offer a substantive interpretation of the CETT defined in their Equation (23) in the different context of social mobility research.

We make three brief remarks about the conditional effect defined in (S6) above. First, discrimination ($X$) remains the focal predictor of interest. The CETT is used merely to investigate whether the effect of $X$ differs across individuals with different values of $Z$. Second, there may be multiple covariates in $Z$, so each stratum is a vector-valued combination. Third, the conditional effects are causal estimands defined conceptually in terms of potential outcomes – without relying on assuming specific statistical interactions between discrimination $X$ and the covariate(s) $Z$ in the outcome model. As we will see next, this has the benefit of admitting estimators which do not rely on correctly assuming the parametric form of the outcome model.

## D.1 Estimating CETTs

We now describe how to estimate the CETTs in different subgroups. The estimator $\widehat{\text{ETT}}_{\text{DR}}$ in (10) can be readily adapted to calculate subgroup-specific effects. After carrying out steps A1-A4 as described above, calculate the following subgroup-specific estimator in step A5:

$$\widehat{\text{CETT}_{\text{DR}}}(z) = \left( \sum_i X_i \mathbb{1}(Z_i = z) \right)^{-1} \sum_i \mathbb{1}(Z_i = z) \{X_i - (1 - X_i)\} W_i D_i. \tag{S7}$$

The indicator function is denoted by $\mathbb{1}(A)$ and takes value one if the event $A$ occurs and zero otherwise. In other words, the estimator in (S7) is obtained simply by calculating (10) only among individuals with $Z = z$. We emphasize that the same assumed outcome model in (9) can be used without specifying additional interactions between covariate $Z$ and treatment $X$. This is because the outcome model is fitted only to the subgroup with $X = 0$, thus implicitly allowing different relationships between $Y$ and $Z$ based on $X$. Furthermore, the estimator explicates the distinct mechanisms that covariates are used for confounding adjustment versus testing effect heterogeneity. Confounding is addressed by including $C$ as predictors of $X$ in the propensity score model and $Y$ in the outcome model (among those with $X = 0$). In contrast, effect heterogeneity is tested by assessing specific subgroups of $Z$. Routine regression analyses conflate both intentions and can render biased estimates when the assumed parametric form is incorrect.

Extending the doubly robust estimator $\widehat{\text{ETT}}_{\text{WLS}}$ in (11) to estimate subgroup-specific effects is more complex. Because the effects are parameterized as coefficients in the assumed outcome model, it is necessary to use a more complex model with interaction(s) between $X$ and $Z$ so that their coefficients parametrize the corresponding CETTs. For example, interest may be in the conditional effects of discrimination due to gender among different age groups in the study. For simplicity, let the age group be an ordinal categorical variable with three levels: younger, middle-aged, or older adults. Then $Z$ could be defined as auxiliary (dummy-coded) variables $Z = (Z_m, Z_o)$, where $Z_m = 1$ if an individual is middle-aged or zero otherwise, and $Z_o = 1$ if an individual is an older adult or zero otherwise. Therefore, there are three unique values of $Z$: (0,0), (1,0), and (0,1). Then, after carrying out step B1 as described above, fit the following model in place of (11):

$$\text{E}(Y|C, X) = \beta_0 + \beta_1 X + \beta_2 X Z_m + \beta_3 X Z_o + \beta_4 C + \beta_5 C^2. \tag{S8}$$

The weighted least squares estimators of the coefficients $(\beta_1, \beta_2, \beta_3)$ in (S8) can then be combined to construct estimators of the CETTs for the different subgroups. Specifically, the estimators are: among middle-aged individuals $\widehat{\text{CETT}}_{\text{WLS}}(1, 0) = \hat{\beta}_1 + \hat{\beta}_2$; among older adults $\widehat{\text{CETT}}_{\text{WLS}}(0, 1) = \hat{\beta}_1 + \hat{\beta}_3$; and among younger individuals $\widehat{\text{CETT}}_{\text{WLS}}(0, 0) = \hat{\beta}_1$.

# E Simulation studies

We conducted Monte Carlo simulation studies to empirically probe the operating characteristics of the doubly robust ETT estimators introduced above, alongside several other estimators which are not doubly robust. In the first study, we demonstrated causal conditions under which the ATE could not be consistently estimated – even with a correctly

specified regression model – but the ETT could be. In the second study, we demonstrated how the doubly robust estimators of the ETT were protected from biases due to one of the regression models being incorrectly specified.

### E.1 Simulation Study 1

The first study aimed to empirically demonstrate how the ETT could be consistently estimated under weaker causal assumptions than the ATE. Specifically, weak ignorability in (5) was satisfied so that estimates of the ETT could be unbiased, but strong ignorability in (4) was violated, thus ruling out consistent estimates of the ATE. Each observed dataset was generated using the following steps:

$$\begin{aligned} U^* &\sim \text{Bernoulli}(0.25) \\ U &= -U^* \\ C &\sim \text{Bernoulli}(0.1) \\ Y^0 &\sim \mathcal{N}(1 - C, 1) \\ Y^1 &= Y^0 + U. \end{aligned}$$

The variable $U$ encoded stochastic individual-level causal effects of discrimination, where 25% of the population experienced a negative unit effect while the remaining 75% experienced a zero effect (so that the actual value of the ATE was $-0.25$). This variable $U$ was hidden from the observed data and could not be adjusted for. The potential outcomes $Y^0$ were independent of $U$, thus ensuring that weak ignorability in (5) would be satisfied. However, the potential outcomes $Y^1$ depended on $U$ because $U$ encoded the individual-level impact of experiencing discrimination $(Y^1 - Y^0)$.

The propensity score and observed outcomes were generated as follows:

$$\begin{aligned} X^* &= -1 + 1.7C + \xi U \\ X &\sim \text{Bernoulli}\left\{\frac{\exp(X^*)}{1 + \exp(X^*)}\right\} \\ Y &= XY^1 + (1 - X)Y^0. \end{aligned}$$

Crucially, individuals with a negative effect ($U = -1$) could choose to avoid experiencing discrimination ($X = 1$), with the strength of the association between $X$ and $U$ parameterized by $\xi$. Hence, when $\xi \neq 0$, $Y^1$ was associated with $X$ due to hidden $U$, violating strong ignorability in (4) and ruling out consistent estimation of the ATE.[2]

One of our goals was to demonstrate the empirical biases resulting from incorrectly specified models for the propensity score or the outcome. Therefore, in addition to the

[2]This data-generating mechanism was a simplified version of that in Morgan and Winship (2015, pages 172-173).

doubly robust estimators – $\widehat{\text{ETT}}_{\text{DR}}$ in (10) and $\widehat{\text{ETT}}_{\text{WLS}}$ using (11) – introduced above, we considered the following estimators that relied on only an outcome model, or a propensity score model, but not both.

- An outcome-only regression estimator:

$$\widehat{\text{ETT}}_{\text{Y}} = \left(\sum_i X_i\right)^{-1} \sum_i X_i D_i. \tag{S9}$$

  This estimator is a simplified variant of the doubly robust estimator in (10) because it uses only an outcome model (hence the subscript Y) for the $X = 0$ subgroup such as in (9). It ignores the weights $W$, thus avoiding assuming a propensity score model for $X$ given $C$. We elaborate on how this estimator was derived in the Appendix.

- An IPW estimator:

$$\widehat{\text{ETT}}_{\text{IPW}} = \left(\sum_i X_i\right)^{-1} \sum_i \{X_i - (1 - X_i)\} W_i Y_i. \tag{S10}$$

  This estimator is a different simplified variant of the doubly robust estimator in (10) because it uses only a propensity score model (to calculate the weights $W$). It utilizes only the raw outcomes $Y$ instead of the computed differences $D$, thus avoiding assuming an outcome model for the associations between $Y$ and $C$. We elaborate on how this estimator was derived in the Appendix.

- Another IPW estimator:

$$\widehat{\text{ETT}}_{\text{Hajek}} = \frac{\sum_i X_i W_i Y_i}{\sum_i X_i W_i} - \frac{\sum_i (1 - X_i) W_i Y_i}{\sum_i (1 - X_i) W_i}. \tag{S11}$$

  This estimator is a simplified variant of the weighted least squares estimator $\widehat{\text{ETT}}_{\text{WLS}}$ using (11) because it uses only a propensity score model (to calculate the weights $W$). It is equivalent to a weighted least squares estimator of the coefficient of $X$ when regressing the raw outcomes $Y$ on $X$ alone (i.e., without the covariates $C$) using the weights $W$ (Morgan & Winship, 2015, page 232). Like the other IPW estimator, no outcome model for the associations between $Y$ and $C$ is assumed. This estimator is more generally termed a Hajek or modified Horvitz-Thompson estimator; see Reifeis and Hudgens (2022, Equation (1)).

We considered sample sizes of $N \in \{200, 5000, 80000\}$ to demonstrate that the biases persist even in large samples, and to compare the relative performance of the different estimators (under correctly specified models) in small samples. We considered $\xi \in \{0.0, 0.7, 1.4, 2.1, 2.8, 3.5\}$ to illustrate the biases as the strength of the association

between $X$ and $U$ increased. We used equally-sized increments of magnitude 0.7 so that the odds of $X = 1$ approximately double (because $\exp(0.7) \approx 2$) in the data-generating process as $\xi$ increases.[3] All results were based on 10000 simulated samples. The results are shown in Table S1. As expected, the estimates of the ATE were unbiased only when $X$ did not depend on $U$ ($\xi = 0$); in all other settings, the estimates of the ATE were biased. In contrast, all the estimators of the ETT were consistent or unbiased. Moreover, the weighted least squares estimator $\widehat{\text{ETT}}_{\text{WLS}}$ using (11) displayed finite sample biases that were reduced only with larger sample sizes.

Table S1: Empirical biases of different estimators of the causal effects in Simulation Study 1.

| | | True | $\widehat{\text{ATE}}$ | | True | $\widehat{\text{ETT}}_{\text{Y}}$ | | $\widehat{\text{ETT}}_{\text{IPW}}$ | | $\widehat{\text{ETT}}_{\text{Hajek}}$ | | $\widehat{\text{ETT}}_{\text{WLS}}$ | | $\widehat{\text{ETT}}_{\text{DR}}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | $\xi$ | ATE | Est. | Bias | ETT | Est. | Bias | Est. | Bias | Est. | Bias | Est. | Bias | Est. | Bias |
| 200 | 0.0 | -0.25 | -0.25 | -0.00 | -0.25 | -0.25 | -0.00 | -0.25 | -0.00 | -0.25 | -0.00 | -0.28 | -0.03 | -0.25 | -0.00 |
| 5000 | 0.0 | -0.25 | -0.25 | -0.00 | -0.25 | -0.25 | -0.00 | -0.25 | -0.00 | -0.25 | -0.00 | -0.25 | -0.00 | -0.25 | -0.00 |
| 80000 | 0.0 | -0.25 | -0.25 | -0.00 | -0.25 | -0.25 | -0.00 | -0.25 | -0.00 | -0.25 | -0.00 | -0.25 | -0.00 | -0.25 | -0.00 |
| 200 | 0.7 | -0.25 | -0.17 | 0.08 | -0.17 | -0.17 | -0.00 | -0.17 | -0.00 | -0.17 | -0.00 | -0.21 | -0.04 | -0.17 | -0.00 |
| 5000 | 0.7 | -0.25 | -0.17 | 0.09 | -0.17 | -0.17 | 0.00 | -0.17 | 0.00 | -0.17 | 0.00 | -0.18 | -0.01 | -0.17 | 0.00 |
| 80000 | 0.7 | -0.25 | -0.17 | 0.08 | -0.17 | -0.17 | -0.00 | -0.17 | -0.00 | -0.17 | -0.00 | -0.18 | -0.01 | -0.17 | -0.00 |
| 200 | 1.4 | -0.25 | -0.10 | 0.15 | -0.10 | -0.10 | -0.00 | -0.10 | -0.00 | -0.10 | -0.00 | -0.15 | -0.04 | -0.10 | -0.00 |
| 5000 | 1.4 | -0.25 | -0.10 | 0.15 | -0.10 | -0.10 | -0.00 | -0.10 | -0.00 | -0.10 | -0.00 | -0.12 | -0.02 | -0.10 | -0.00 |
| 80000 | 1.4 | -0.25 | -0.10 | 0.15 | -0.10 | -0.10 | 0.00 | -0.10 | 0.00 | -0.10 | 0.00 | -0.12 | -0.02 | -0.10 | 0.00 |
| 200 | 2.1 | -0.25 | -0.05 | 0.20 | -0.06 | -0.06 | 0.00 | -0.06 | 0.00 | -0.06 | 0.00 | -0.11 | -0.05 | -0.06 | 0.00 |
| 5000 | 2.1 | -0.25 | -0.06 | 0.19 | -0.06 | -0.06 | 0.00 | -0.06 | 0.00 | -0.06 | 0.00 | -0.08 | -0.02 | -0.06 | 0.00 |
| 80000 | 2.1 | -0.25 | -0.06 | 0.19 | -0.06 | -0.06 | -0.00 | -0.06 | -0.00 | -0.06 | -0.00 | -0.08 | -0.02 | -0.06 | -0.00 |
| 200 | 2.8 | -0.25 | -0.03 | 0.22 | -0.03 | -0.03 | 0.00 | -0.03 | 0.00 | -0.03 | 0.00 | 0.07 | 0.11 | -0.03 | 0.00 |
| 5000 | 2.8 | -0.25 | -0.03 | 0.22 | -0.03 | -0.03 | -0.00 | -0.03 | -0.00 | -0.03 | -0.00 | -0.05 | -0.02 | -0.03 | -0.00 |
| 80000 | 2.8 | -0.25 | -0.03 | 0.22 | -0.03 | -0.03 | -0.00 | -0.03 | -0.00 | -0.03 | -0.00 | -0.05 | -0.02 | -0.03 | -0.00 |
| 200 | 3.5 | -0.25 | -0.01 | 0.24 | -0.02 | -0.02 | 0.00 | -0.02 | 0.00 | -0.02 | 0.00 | 0.00 | 0.02 | -0.02 | 0.00 |
| 5000 | 3.5 | -0.25 | -0.01 | 0.23 | -0.02 | -0.02 | 0.00 | -0.02 | 0.00 | -0.02 | 0.00 | -0.03 | -0.01 | -0.02 | 0.00 |
| 80000 | 3.5 | -0.25 | -0.01 | 0.23 | -0.02 | -0.02 | 0.00 | -0.02 | 0.00 | -0.02 | 0.00 | -0.03 | -0.01 | -0.02 | 0.00 |

*Note.* $N$=sample size; ATE=average treatment effect; ETT=effect of treatment on the treated; IPW=Inverse probability of treatment weighting; WLS=weighted least squares; DR=doubly robust; "Est."=Estimate. The true value of the ETT was calculated in each simulated sample as the average difference of the potential outcomes among those who experienced discrimination ($X = 1$). The estimator of the ATE ($\widehat{\text{ATE}}$) was the ordinary least squares estimator of the coefficient of $X$ in a (correctly-specified) regression of $Y$ on $X$ and $C$, such as in (7). The estimators of the ETT utilized a correctly assumed model for the propensity score or the outcome. The columns displaying the biases were shaded grey to improve readability. All results were rounded to two decimal places.

[3] We thank anonymous reviewer 2 for suggesting this clarification.

## E.2 Simulation Study 2

The second study aimed to empirically demonstrate biases brought on by incorrectly specified models for either the propensity score or the outcome, and to compare the reduced biases when utilizing the doubly robust estimators. Each observed dataset was generated using the following steps:

$$
\begin{aligned}
C_1 &\sim \text{Uniform}(-\sqrt{3}, \sqrt{3}) \\
C_2 &\sim \text{Bernoulli}(0.1) \\
U &= -0.6 + \epsilon_U, \quad \epsilon_U \sim \text{Uniform}\{-0.3, 0.3\} \\
Y^0 &= \mu^0(C) + \epsilon_Y, \quad \epsilon_Y \sim \mathcal{N}(0,1) \\
Y^1 &= Y^0 + U \\
X^* &= \text{Prob}(X = 1|C, U) \\
X &\sim \text{Bernoulli}\left\{\frac{\exp(X^*)}{1+\exp(X^*)}\right\} \\
Y &= XY^1 + (1-X)Y^0.
\end{aligned}
$$

The baseline covariates comprised of two variables: $C = (C_1, C_2)$. The mean for the potential outcome $Y^0$, denoted by $\mu^0(C)$, was a function of $C_1$ and $C_2$. We considered two possibilities for $\mu^0(C)$:

1. $\mu^0(C) = C_1 - C_1C_2$; or
2. $\mu^0(C) = \sqrt{|C_1 + 0.5|}$.

The latter was a nonlinear function of $C_1$ intended to induce model misspecification biases when using routine linear models for the outcome (even after including an interaction between $C_1$ and $C_2$). The propensity score, represented by $\text{Prob}(X = 1|C, U)$, was a function of $C$ and $U$. We considered two possibilities for $\text{Prob}(X = 1|C, U)$:

1. $\text{Prob}(X = 1|C, U) = C_1 + C_1C_2 - 6U$; or
2. $\text{Prob}(X = 1|C, U) = |C_1|(3 - C_1^2) + 2(C_2 - 2)\min(|C_1|^{2C_1}, 2) - 6U$.

The latter was a complex nonlinear function of the covariates in $C$ intended to induce model misspecification biases when using routine log-linear logistic regression models for the propensity score (even after including an interaction term between $C_1$ and $C_2$). The variable $U$ encoded stochastic individual-level causal effects, uniformly distributed between $-0.9$ and $-0.3$, that were hidden from the observed data to induce unmeasured confounding between $Y^1$ and $X$. Hence, the ATE equaled $-0.6$ (the average of $U$).

We considered sample sizes of $N \in \{200, 5000, 80000\}$ to demonstrate that model misspecification biases persist even in large samples. All results were based on 10000 simulated

samples. The results are shown in Table S2. The ETT estimators which relied on a single model were unbiased only when that model was correctly specified. In contrast, the doubly robust estimators were consistent in all scenarios, except when both models were incorrectly specified. When both models were incorrectly specified, the biases using the doubly robust estimators were similar to, or less biased than, the other estimators.

Table S2: Empirical biases of different estimators of the causal effects in Simulation Study 2.

| | True model | | True | $\widehat{\text{ATE}}$ | | True | $\widehat{\text{ETT}}_{\text{Y}}$ | | $\widehat{\text{ETT}}_{\text{IPW}}$ | | $\widehat{\text{ETT}}_{\text{Hajek}}$ | | $\widehat{\text{ETT}}_{\text{WLS}}$ | | $\widehat{\text{ETT}}_{\text{DR}}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | PS | Outcome | ATE | Est. | Bias | ETT | Est. | Bias | Est. | Bias | Est. | Bias | Est. | Bias | Est. | Bias |
| 200 | linear | linear | -0.60 | -0.62 | -0.02 | -0.61 | -0.51 | 0.09 | -0.50 | 0.11 | -0.42 | 0.19 | -0.60 | 0.00 | -0.51 | 0.10 |
| 5000 | linear | linear | -0.60 | -0.62 | -0.02 | -0.61 | -0.61 | 0.00 | -0.59 | 0.02 | -0.59 | 0.02 | -0.61 | 0.00 | -0.61 | 0.00 |
| 80000 | linear | linear | -0.60 | -0.62 | -0.02 | -0.61 | -0.61 | -0.00 | -0.60 | 0.01 | -0.60 | 0.01 | -0.61 | -0.00 | -0.61 | -0.00 |
| 200 | nonlinear | linear | -0.60 | -0.63 | -0.03 | -0.63 | -0.59 | 0.04 | -0.71 | -0.08 | -0.71 | -0.08 | -0.63 | -0.00 | -0.59 | 0.04 |
| 5000 | nonlinear | linear | -0.60 | -0.63 | -0.03 | -0.63 | -0.63 | -0.00 | -0.71 | -0.08 | -0.70 | -0.08 | -0.63 | -0.00 | -0.63 | -0.00 |
| 80000 | nonlinear | linear | -0.60 | -0.63 | -0.03 | -0.63 | -0.63 | -0.00 | -0.70 | -0.08 | -0.70 | -0.08 | -0.63 | -0.00 | -0.63 | -0.00 |
| 200 | linear | nonlinear | -0.60 | -0.71 | -0.11 | -0.61 | -0.39 | 0.22 | -0.46 | 0.15 | -0.53 | 0.08 | -0.57 | 0.04 | -0.44 | 0.17 |
| 5000 | linear | nonlinear | -0.60 | -0.70 | -0.10 | -0.61 | -0.47 | 0.14 | -0.58 | 0.02 | -0.60 | 0.01 | -0.60 | 0.01 | -0.58 | 0.03 |
| 80000 | linear | nonlinear | -0.60 | -0.70 | -0.10 | -0.61 | -0.47 | 0.14 | -0.59 | 0.02 | -0.60 | 0.01 | -0.60 | 0.00 | -0.59 | 0.01 |
| 200 | nonlinear | nonlinear | -0.60 | -0.41 | 0.19 | -0.63 | -0.57 | 0.06 | -0.33 | 0.30 | -0.36 | 0.27 | -0.34 | 0.29 | -0.58 | 0.04 |
| 5000 | nonlinear | nonlinear | -0.60 | -0.41 | 0.19 | -0.63 | -0.32 | 0.31 | -0.36 | 0.27 | -0.37 | 0.26 | -0.35 | 0.28 | -0.33 | 0.30 |
| 80000 | nonlinear | nonlinear | -0.60 | -0.41 | 0.19 | -0.63 | -0.32 | 0.31 | -0.36 | 0.27 | -0.37 | 0.26 | -0.35 | 0.28 | -0.33 | 0.30 |

*Note.* $N$=sample size; PS=propensity score; ATE=average treatment effect; ETT=effect of treatment on the treated; IPW=Inverse probability weighting; WLS=weighted least squares; DR=doubly robust; "Est."=Estimate. The true value of the ETT was calculated in each simulated sample as the average difference of the potential outcomes among those who experienced discrimination ($X = 1$). The estimator of the ATE ($\widehat{\text{ATE}}$) was the ordinary least squares estimator of $\beta_1$ in the outcome model $\text{E}(Y|X, C) = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \beta_4 C_1 C_2 + \beta_5 C_1^2$. This same outcome model was also used to calculate the estimator $\widehat{\text{ETT}}_{\text{WLS}}$. For the other estimators, we assumed in the PS and outcome models the main effects for $C_1$ and $C_2$, an interaction term $C_1 C_2$, and a quadratic term for $C_1$. Because the assumed PS and outcome models for constructing the estimators were linear, an assumed model was (in)correctly specified when the true model was (non)linear. The columns displaying the biases were shaded grey to improve readability. All results were rounded to two decimal places.

## References


Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, *18*(1), 36–56. https://doi.org/10.1093/pan/mpp036

Moodie, E. E. M., Saarela, O., & Stephens, D. A. (2018). A doubly robust weighting estimator of the average treatment effect on the treated. *Stat*, *7*(1), e205. https://doi.org/10.1002/sta4.205

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Reifeis, S. A., & Hudgens, M. G. (2022). On variance of the treatment effect in the treated when estimated by inverse probability weighting. *American Journal of Epidemiology*, *191*(6), 1092–1097. https://doi.org/10.1093/aje/kwac014

Robins, J. M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, *22*(4), 544–559. https://doi.org/10.1214/07-STS227D

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313. https://doi.org/10.1037/a0014268

Zang, E., Sobel, M. E., & Luo, L. (2023). The mobility effects hypothesis: Methods and applications. *Social Science Research*, *110*, 102818. https://doi.org/10.1016/j.ssresearch.2022.102818