Received: 20 May 2019 | Revised: 22 February 2020

DOI: 10.1002/jae.2765

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Direct and indirect effects of continuous treatments based on generalized propensity score weighting

Martin Huber[1] | **Yu-Chin Hsu**[2] | **Ying-Ying Lee**[3] | **Layal Lettry**[4]

[1]Department of Economics, University of Fribourg, Fribourg, Switzerland

[2]Academia Sinica, Institute of Economics, National Central University, Taoyuan City, Taiwan

[3]Department of Economics, University of California Irvine, Irvine, California

[4]Swiss Federal Agency for Social Insurances, Bern, Switzerland

Correspondence
Martin Huber, Department of Economics, University of Fribourg, Bd de Pérolles 90, 1700 Fribourg, Switzerland.
Email: martin.huber@unifr.ch

Summary

This paper proposes semi- and nonparametric methods for disentangling the total causal effect of a continuous treatment on an outcome variable into its natural direct effect and the indirect effect that operates through one or several intermediate variables called mediators jointly. Our approach is based on weighting observations by the inverse of two versions of the generalized propensity score (GPS), namely the conditional density of treatment either given observed covariates or given covariates and the mediator. Our effect estimators are shown to be asymptotically normal when the GPS is estimated by either a parametric or a nonparametric kernel-based method. We also provide a simulation study and an empirical illustration based on the Job Corps experimental study.

1 | INTRODUCTION

Classic treatment evaluations typically focus on assessing the total causal effect of a treatment on an outcome variable—for example, the average treatment effect (ATE). In many evaluation problems, however, the causal mechanisms through which a total effect operates are also of interest. When, for example, assessing the effect of an educational program on criminal activity, policymakers might want to learn whether the total effect is driven by the program's effect on employment chances, which in turn may affect criminal behavior, or by other features of the program such as its impact on personality traits like integrity or discipline. Understanding the causal mechanisms may be helpful for appropriately designing such educational programs—for example, whether the focus should be on increasing employability, personality development, or both.

Causal mediation analysis aims to decompose a total treatment effect into the indirect effect operating through an intermediate variable called mediator, and the direct effect net of mediation; see, for instance, Robins and Greenland (1992) and Pearl (2001). A range of studies base identification on conditional independence assumptions given observables with respect to treatment and mediator assignment in rather flexible (often nonparametric) models; see, for instance, Petersen, Sinisi, and van der Laan (2006), Flores and Flores-Lagunes (2009), van der Weele (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Tchetgen Tchetgen and Shpitser (2012), and Vansteelandt, Bekaert, and Lange (2012), among others.[1] Contributions concerned with nonparametric identification under conditional independence conventionally focus on binary treatments. Yet, there are many empirical problems in which treatment intensity is (close to) continuous—for example, hours of participation in an educational program or the dose of a medical treatment; see, for instance, Hirano and Imbens (2004), Imai and van Dyk (2004), Bia and Mattei (2012), Flores, Flores-Lagunes, Gonzalez, and Neumann (2012), Kluve, Schneider, Uhlendorff, and Zhao (2012), Galvao and Wang (2015), and Lee (2018).

[1]In contrast, the seminal papers in mediation analysis of Judd and Kenny (1981) and Baron and Kenny (1986) assume linear models for both the mediator and the outcome.

This paper considers the identification and semi- as well as nonparametric estimation of natural direct and indirect effect (in the denomination of Pearl, 2001)[2] when the treatment is continuous. The indirect effect might either concern a single mediator or reflect the impact operating through multiple mediators jointly. In the latter case, conditional independence must hold for each mediator. The joint indirect effect then contains the causal mechanisms working through any of these mediators (possibly including interaction effects between mediators), while the direct effect is the remainder impact net of any of these causal mechanisms; see VanderWeele and Vansteelandt (2014) and Huber (2019) for a more thorough discussion of the multiple mediators framework.[3] We propose an estimator based on weighting by the inverse of conditional treatment densities (i) given observed covariates and (ii) given covariates and the mediator(s), also known as generalized propensity scores; see Hirano and Imbens (2004) and Imai and van Dyk (2004).

The generalized propensity scores are either obtained parametrically or nonparametrically by conditional kernel density estimation. We show that estimation is asymptotically normal and converges at the rate of one-dimensional nonparametric regression to the effects of interest under specific regularity conditions. We also provide a simulation study that illustrates the robustness of our method when compared to classic linear mediation analysis that relies on tight parametric assumptions. Finally, we apply our approach to data on the Job Corps program, a US educational intervention for disadvantaged youth. Specifically, we disentangle the negative effect of the length of exposure to academic and vocational instruction in Job Corps on crime, measured by the number of arrests in the fourth year, into an indirect component operating through the mediator employment and a direct remainder effect. The latter covers any other causal mechanisms, such as personality development. Our findings point to an important direct and nonlinear reduction of the number of arrests as a consequence of Job Corp under a sufficiently large treatment intensity of roughly 1,000 hours or more, while indirect effects are close to zero for the investigated range of treatment intensities of up to 2,000 hours.

Our paper fills an important methodological gap in the causal mediation literature with continuous treatment doses, where studies typically rely on rather strong functional form restrictions for identification. The semi- and nonparametric literature on continuous treatments under conditional independence is relatively sparse and focuses on the estimation of total (rather than direct and indirect) treatment effects: Flores (2007) proposes a nonparametric kernel regression estimator for average dose–response functions. Lee (2018) estimates the unconditional distribution of potential outcomes using the estimated generalized propensity score as generated regressor. Flores (2005), Flores et al. (2012), and Galvao and Wang (2015) discuss estimation based on weighting by the inverse of the generalized propensity score. Our approach can be regarded as an extension of the semi- and nonparametric weighting approaches of Huber (2014) and Hsu, Huber, and Lai (2018) for causal mediation analysis with discrete treatments to the continuous treatment case using kernel functions and the concept of the generalized propensity score. The semiparametric version of the proposed estimator is available in the "causalweight" package by Bodory and Huber (2018) for the statistical software "R."[4]

The remainder of the paper is organized as follows. Section 2 introduces the parameters of interest. Section 3 discusses the identifying assumption and identification based on weighting. Section 4 presents the estimation approach along with its properties. Sections 5 and 6 provide a simulation study and empirical illustration based on the Job Corps experimental study, respectively. Section 7 concludes.

2 | PARAMETERS OF INTEREST

Our goal is to decompose the average treatment effect (ATE) of a continuous treatment variable D on an outcome variable Y into a direct effect and an indirect effect operating through the mediator M, which may be a scalar or a vector and discrete and/or continuous. For a generic random variable A, let $\mathcal{A}$ denote the support of A. To define the effects of interest, we use the potential outcome framework (e.g., Rubin, 1974), which has been applied in the context of mediation analysis by Rubin (2004), Ten Have et al. (2007), and Albert (2008), among others. Let $M(d)$, $Y(d, M(d'))$ denote the potential mediator state as a function of the treatment and potential outcome as a function of the treatment and the potential mediator, respectively, under treatments values $d, d' \in \mathcal{D}$. Furthermore, denote the mean potential outcomes by $\mu(d, d) = E[Y(d, M(d))]$ and

[2]Such effects have also been referred to as pure/total direct and indirect effects by Robins and Greenland (1992) and Robins (2003) or as net and mechanism treatment effects by Flores and Flores-Lagunes (2009).

[3]The joint indirect effect of multiple mediators generally differs from the sum of the indirect effects when considering each mediator separately (even when appropriately accounting for interaction effects between mediators), unless statistical associations across mediators are ruled out; see Imai and Yamamoto (2013).

[4]Further alternatives for assessing direct and indirect effects of continuous treatments are the "medflex" package by Steen, Loeys, Moerkerke, and Vansteelandt (2017), which implements imputation-based estimation of potential outcomes as suggested by Vansteelandt et al. (2012), and the regression-based "mediation" package by Tingley, Yamamoto, Hirose, Imai, and Keele (2014).

$\mu(d, d') = E[Y(d, M(d'))]$ with $d \neq d'$. We note that under a continuous treatment $\mu(d, d)$ has also been referred to as average dose–response function in the literature; see, for instance, Hirano and Imbens (2004) and Imai and van Dyk (2004).

Using this notation, the ATE of setting the treatment to d versus d', denoted by $\Delta_{d,d'}$, is given by

$$\Delta_{d,d'} = \mu(d, d) - \mu(d', d'), \quad \text{for} \quad d \neq d'. \tag{1}$$

The ATE thus corresponds to the total average effect of D on Y operating both indirectly via the difference in potential mediators $M(d)$ and $M(d')$ as well as directly. In contrast, the average natural direct effect is given by the difference in mean potential outcomes under d versus d' when keeping the potential mediator fixed at either $M(d)$ or $M(d')$:

$$\theta_{d,d'}(d') = \mu(d, d') - \mu(d', d'), \quad \theta_{d,d'}(d) = \mu(d, d) - \mu(d', d), \quad \text{for} \quad d \neq d'. \tag{2}$$

Analogously, the average natural indirect effect is defined as the difference in mean potential outcomes under $M(d)$ versus $M(d')$ while keeping the treatment fixed at either d or d' such that the direct effect is nil:

$$\delta_{d,d'}(d) = \mu(d, d) - \mu(d, d'), \quad \delta_{d,d'}(d') = \mu(d', d) - \mu(d', d'), \quad \text{for } d \neq d'. \tag{3}$$

By adding and subtracting $\mu(d, d')$ in Equation 1, it is easy to verify that $\theta_{d,d'}(d')$ and $\delta_{d,d'}(d)$ add up to the ATE. By similarly adding and subtracting $\mu(d', d)$ in Equation 1, one sees that the ATE also corresponds to the sum of $\theta_{d,d'}(d)$ and $\delta_{d,d'}(d')$, where d and d' in the definition of direct and indirect effects have been swapped. That is, $\Delta_{d,d'} = \theta_{d,d'}(d') + \delta_{d,d'}(d) = \theta_{d,d'}(d) + \delta_{d,d'}(d')$. Indeed, $\theta_{d,d'}(d')$ and $\delta_{d,d'}(d)$ might differ from $\theta_{d,d'}(d)$ and $\delta_{d,d'}(d')$, respectively, if direct and indirect effects are heterogeneous in M and D, respectively. This is the case in the presence of interactions of D and M in the determination of outcome Y.

We note that if $d - d' \to 0$ such that the change in D becomes infinitesimal, $\lim_{d' \to d}(\mu(d, d) - \mu(d', d'))/(d - d')$ corresponds to the derivative of $\mu(d, d)$ w.r.t. d, denoted by $\frac{\boldsymbol{d}\mu(d,d)}{\boldsymbol{d}d}$. This (total) marginal effect of the treatment at $D = d$ has, for instance, been considered in Hirano and Imbens (2004) and Flores et al. (2012). The total derivative can be written as

$$\frac{\boldsymbol{d}}{\boldsymbol{d}t}\mu(t, t)\Big|_{t=d} = \frac{\partial}{\partial t}\mu(t, d) + \frac{\partial}{\partial t}\mu(d, t)\Big|_{t=d}, \tag{4}$$

where $\boldsymbol{d}$ and ∂ denote total and partial derivatives, respectively. $\frac{\partial}{\partial t}\mu(t, d)|_{t=d}$ and $\frac{\partial}{\partial t}\mu(d, t)|_{t=d}$ are the marginal direct and indirect effects, respectively, and correspond to Equations 2 and 3 divided by $d - d'$ when letting $d - d' \to 0$.[5] Even though our assumptions presented in Section 3 also permit identifying the marginal total, direct, and indirect effects provided in Equation 4, the discussion in this paper focuses on the effects under measurable treatment changes such that $d \neq d'$. This permits investigating effect heterogeneities due to interactions of D and M by comparing $\theta_{d,d'}(d)$ and $\theta_{d,d'}(d')$ as well as $\delta_{d,d'}(d)$ and $\delta_{d,d'}(d')$, respectively, while such interactions are conceptually ruled out under infinitesimal changes in D.

3 | IDENTIFICATION

For each unit only one potential outcome and potential mediator state, respectively, are known, namely those related to the treatment value that is observed for that unit. That is, the observed mediator and outcome correspond to $M = M(D)$ and $Y = Y(D, M(D))$ under the observed treatment state D. In contrast, we cannot observe potential outcomes and mediators defined upon treatment values different from the observed one. Specifically, $Y(d, M(d'))$ is not observed for any individual if $d \neq d'$, as at least one of d, d' is necessarily different to the observed treatment.

[5] Suppose that $Y(d, M(d))$ is differentiable in both arguments, d and $M(d)$, and $M(d)$ is a scalar and is differentiable in d. Then the marginal indirect effect can be written as

$$\frac{\partial}{\partial t}\mu(d, t)\Big|_{t=d} = E\left[\frac{\partial Y(d, M(d))}{\partial M(d)}\frac{\boldsymbol{d}M(d)}{\boldsymbol{d}d}\right].$$

This is equivalent to the indirect effect in linear models (first-stage effect, $\frac{\boldsymbol{d}M(d)}{\boldsymbol{d}d}$, times second-stage effect, $\frac{\partial Y(d,M(d))}{\partial M(d)}$) when there are no D–M interactions (see, e.g., Baron & Kenny, 1986).

The identification of natural direct and indirect effects therefore requires specific assumptions. Similar to Imai, Keele, and Yamamoto (2010) (see their assumption 1), Tchetgen Tchetgen and Shpitser (2012) and many others, we base identification on a sequential conditional independence assumption imposed on treatment and mediator assignment. However, contrary to the standard in the literature, we consider a continuous treatment rather than a binary one.

Our first assumption requires that, given a vector of observed pretreatment characteristics that we denote by X, the treatment is conditionally independent of the potential mediator states and the potential outcomes.

Assumption 1 (Conditional independence of the treatment)**.** $\{Y(d', m), M(d)\} \perp D|X = x$ for all $(d, d', m, x) \in \mathcal{D}^2 \times \mathcal{M} \times \mathcal{X}$.

Assumption 1 rules out unobserved confounders jointly affecting the treatment, on the one hand, and the mediator and/or the outcome on the other hand, conditional on X. In the treatment or program evaluation literature, this is referred to as conditional independence, selection on observables, or exogeneity; see Imbens (2004). We point out that conditional independence must hold with respect to any value in the continuous support of the treatment, which is stronger than that for the binary treatment case.

Our second assumption imposes conditional independence of the mediator given the treatment and the covariates along with a common support restriction on the conditional density of the treatment. To this end, let $f_A(a|B = b)$ denote the conditional density of variable A at some value a given that variable B is equal to value b.

Assumption 2 (Conditional independence of the mediator)**.**

(i) $Y(d', m) \perp M(d)|D = d, X = x$ for all $(d, d', m, x) \in \mathcal{D}^2 \times \mathcal{M} \times \mathcal{X}$.
(ii) $f_D(d|M = m, X = x) > 0$ for all $(d, m, x) \in \mathcal{D} \times \mathcal{M} \times \mathcal{X}$.

Assumption 2(i) rules out unobserved confounders jointly affecting the mediator and the outcome conditional on D and X. This is for instance violated if unobserved posttreatment variables influence M and Y, and are not fully determined by X and/or D. When M is multidimensional, Assumption 2(i) needs to hold for each element in M, such that its strength increases in the number of mediators. Assumption 2(ii) is a common support restriction. It says that the conditional density (or generalized propensity score) to receive any treatment d in the support of D given M, X is larger than zero. This also implies that $f_D(d|X = x) > 0$ and $f_M(m|D = d, X = x) > 0$ by Bayes' theorem. Intuitively, it is required that individuals (a) with comparable values in M and X exist across all possible treatment doses and (b) with comparable values in D and X exist across all possible mediator values. This common support condition is stronger than that conventionally imposed in the binary treatment case because it needs to hold over the entire support of the treatment, unless only a subset of treatment values d was to be considered in the analysis. Furthermore, it becomes stronger as the number and support of mediators increase.

Huber (2014) shows the identification of the mean potential outcomes $\mu(d, d)$ and $\mu(d, d')$ with $d \neq d'$ using weighting by the inverse of specific propensity scores when Assumptions 1 and 2 are phrased in a binary context. Specifically,

$$\mu(d, d) = E\left[\frac{Y \cdot 1(D = d)}{\Pr(D = d|X)}\right], \tag{5}$$

$$\mu(d, d') = E\left[\frac{Y \cdot 1(D = d)}{\Pr(D = d|M, X)} \cdot \frac{\Pr(D = d'|M, X)}{\Pr(D = d'|X)}\right], \tag{6}$$

$1(\cdot)$ denoting the indicator function. Also, $\Pr(D = d|X) = E[1(D = d)|X]$ and $\Pr(D = d|M, X) = E[1(D = d)|M, X]$ are the conditional expectations of the weights, $1(D = d)$, that correspond to the treatment propensity scores. In the binary treatment case, Equations 5 and 6 therefore correspond to equations 4 and 5 in Huber (2014).

Closely related identification results can be established for the case of a continuous treatment when appropriately adapting the weighting expressions; see, for instance, the discussion in Flores et al. (2012) and Flores (2005). To this end, denote by $\omega(D; d, h)$ a weighting function that depends on the distance between D and the reference value d as well as a nonnegative tuning parameter h. The closer the tuning parameter h is to zero, the less weight is given to larger discrepancies between D and d. This modification of the weighting function is required as truly continuous treatments do not have mass points. The probability of a specific value d is therefore equal to zero, which excludes the use of indicator functions. For example, as in Flores et al. (2012), we define the weighting function to be a kernel function: $\omega(D; d, h) \equiv K((D - d)/h)/h$, where K is a symmetric second-order kernel function assigning more weight to observations closer to d and h is a bandwidth. Under the assumption that $f_D(d|M, X)$ and $E[Y|D = d, M, X]$ are continuous in d, the parameters of

interest are identified in analogy to Equations 5 and 6 when letting h go to zero:

$$\mu(d,d) = \lim_{h\to 0} E\left[\frac{Y\cdot\omega(D;d,h)}{f_D(d|X)}\right], \tag{7}$$

$$\mu(d,d') = \lim_{h\to 0} E\left[\frac{Y\cdot\omega(D;d,h)}{f_D(d|M,X)}\cdot\frac{f_D(d'|M,X)}{f_D(d'|X)}\right], \tag{8}$$

where $f_D(d|X)$ and $f_D(d|M,X)$ are the generalized propensity scores that correspond to $\lim_{h\to 0}E[\omega(D;d',h)|X]$ and $\lim_{h\to 0}E[\omega(D;d',h)|M,X]$, respectively. The identification of the mean potential outcomes implies the identification of the direct and indirect effects defined in Equations 2 and 3.

Finally, we note that identification of mean potential outcomes and effects is alternatively obtained by the following expressions related to the so-called mediation formula—see, for example, Pearl (2001) and Imai, Keele, and Yamamoto (2010):

$$\mu(d,d) = E\left[E[Y|D=d,X=x]\right], \tag{9}$$

$$\begin{aligned}\mu(d,d') &= \int E[Y|D=d,M=m,X=x]dF_{M|D=d',X=x}(m)dF_X(x) \\ &= \int E[Y|D=d,M=m,X=x]\cdot\frac{f_D(d'|M,X)}{f_D(d'|X)}dF_{M|X=x}(m)dF_X(x) \\ &= E\left[E[Y|D=d,M=m,X=x]\cdot\frac{f_D(d'|M,X)}{f_D(d'|X)}\right].\end{aligned} \tag{10}$$

The last two equalities in Equation 10 follow from Bayes' theorem and the law of iterated expectations, respectively. Equations 9 and 10 suggest conducting mediation analysis using nonparametric regression-based estimates of the conditional means $E[Y|D=d,X]$ and $E[Y|D=d,M,X]$, or alternatively of $E[Y|D=d,f_D(d|X)]$ and $E[Y|D=d,f_D(d|M,X)]$, respectively, given the balancing property of the (generalized) propensity score. The balancing property implies $1(D=d)\perp X|f_D(d|X)$ and $1(D=d)\perp\{M,X\}|f_D(d|M,X)$; see Rosenbaum and Rubin (1983) and Hirano and Imbens (2004). Flores et al. (2012) point out that such a regression approach is computationally more burdensome than weighting estimation, in our case based on Equations 7 and 8 as suggested in Section 4, because the respective conditional means need to be computed for each observation in the sample. This is particularly relevant when using the bootstrap for inference, as in our application in Section 6. On the other hand, weighting may be less stable (i.e., prone to a higher variance) than conditional mean regression if $f_D(d|X)$ or $f_D(d|M,X)$ are close to zero; see the discussion in Khan and Tamer (2010).

4 | ESTIMATION

Suppose the availability of a random sample $\{(Y_i,M_i,D_i,X_i)\}_{i=1}^n$ from the joint distribution of (Y,M,D,X) for estimating the potential outcomes as well as the direct and indirect effects. We first describe fully nonparametric estimation of direct and indirect effects based on kernel methods along with its properties. At the end of this section, we discuss semiparametric estimation based on parametric generalized propensity scores. Following standard practice, the subsequent discussion implicitly assumes that regressors have been standardized by dividing by their respective standard deviations.

For an s-dimensional vector $u=(u_{(1)},\ldots,u_{(s)})'$, let $K_h(u)=\Pi_{\ell=1}^s k(u_{(\ell)}/h)/h$ be a product kernel with a generic kernel function k and bandwidth h. Let $K_{1,h_1}(u)=\Pi_{\ell=1}^s k_1(u_{(\ell)}/h_1)/h_1$ and h_1 denote the kernel function and bandwidth, respectively, for the estimation of the generalized propensity scores, and K_{2,h_2} and h_2 be the respective parameters for estimating the mean potential outcomes (based on conditioning only on D). In the first step, the generalized propensity scores—that is, the conditional densities of D given X or M,X—are obtained by

$$\begin{aligned}\widehat{f}_D(d|X_i) &= \frac{\sum_{j=1}^n K_{1,h_1}(X_j-X_i,D_j-d)}{\sum_{j=1}^n K_{1,h_1}(X_j-X_j)}, \\ \widehat{f}_D(d|M_i,X_i) &= \frac{\sum_{j=1}^n K_{1,h_1}(M_j-M_i,X_j-X_i,D_j-d)}{\sum_{j=1}^n K_{1,h_1}(M_j-M_i,X_j-X_j)},\end{aligned} \tag{11}$$

respectively. In the second step, Equations 7 and 8 are estimated by the respective sample analogs with normalized weights, which we denote by $\widehat{\mu}(d,d)$ and $\widehat{\mu}(d,d')$:

$$\widehat{\mu}(d,d) = \sum_{i=1}^{n} \frac{Y_i K_{2,h_2}(D_i - d)}{\widehat{f}_D(d|X_i)} \Bigg/ \sum_{i=1}^{n} \frac{K_{2,h_2}(D_i - d)}{\widehat{f}_D(d|X_i)}, \tag{12}$$

$$\widehat{\mu}(d,d') = \sum_{i=1}^{n} \frac{Y_i K_{2,h_2}(D_i - d)}{\widehat{f}_D(d|M_i,X_i)} \cdot \frac{\widehat{f}_D(d'|M_i,X_i)}{\widehat{f}_D(d'|X_i)} \Bigg/ \sum_{i=1}^{n} \frac{K_{2,h_2}(D_i - d)}{\widehat{f}_D(d|M_i,X_i)} \cdot \frac{\widehat{f}_D(d'|M_i,X_i)}{\widehat{f}_D(d'|X_i)}. \tag{12}$$

Then estimators for natural direct effects $\theta_{d,d'}(d)$ and $\theta_{d,d'}(d')$, and natural indirect effects $\delta_{d,d'}(d)$ and $\delta_{d,d'}(d')$ are given by

$$\widehat{\theta}_{d,d'}(d) = \widehat{\mu}(d,d) - \widehat{\mu}(d',d), \quad \widehat{\theta}_{d,d'}(d') = \widehat{\mu}(d,d') - \widehat{\mu}(d',d'),$$
$$\widehat{\delta}_{d,d'}(d) = \widehat{\mu}(d,d) - \widehat{\mu}(d,d'), \quad \widehat{\delta}_{d,d'}(d') = \widehat{\mu}(d',d) - \widehat{\mu}(d',d').$$

Assumption 3 invokes several regularity conditions required for the consistency and asymptotic normality of the proposed estimator.

Assumption 3 (Regularity conditions)**.**

(i) The data $\{Y_i, M_i, D_i, X_i\}$, $i = 1, \ldots, n$ are independent and identically distributed (i.i.d.).

(ii) The probability density function $f_{DMX}(d,m,x)$ is bounded away from zero and is at least r-order continuously differentiable with respect to (d,m,x), with uniformly bounded derivatives on $\mathcal{D}\times\mathcal{M}\times\mathcal{X}$, a compact and convex subset of $\mathcal{R}^{1+s_m+s_x}$, where s_m and s_x are the dimensions of M and X, respectively.

(iii) $E\left[Y|D=d, M=m, X=x\right]$ is at least r-order continuously differentiable with respect to (d,m,x) on $\mathcal{D}\times\mathcal{M}\times\mathcal{X}$ and has uniformly bounded derivatives.

(iv) The symmetric kernels k_1 and k_2 are bounded differentiable, have convex bounded supports, and have order $r_1 \geq 2$ and $r_2 \geq 2$, respectively.[6]

(v) The bandwidths h_1, h_2 and $h \equiv \min\{h_1, h_2\}$ and the orders r_1 and r_2 satisfy $h_1, h_2 \to 0$, $nh{h_1}^{2s}h_2^2h^{-1} \to \infty$, $nhh_1^{4r_1}h_2^{-2} \to 0$, $nh_1h_2^{2r_2} = O(1)$, $nh_1^{2r_1+1} = O(1)$, $h_1^{2r_1}h_2^{-1}h \to 0$, $nhh_1^{2r_1} \to 0$, and $nhh_2^{2r_2} \to 0$, as $n \to \infty$, where the dimension of the regressors is $s \equiv 1 + s_m + s_x$.

Our estimator can be linearized to follow a U-statistic, which is well studied in the literature. The smoothness and bandwidth conditions in Assumption 3 ensure that the remainder terms of the projections of the U-statistic and the bias terms are asymptotically first-order negligible. Assumption 3(iv) imposes standard regularity conditions for kernel functions. Assumption 3(v) implies that the first step estimators of the conditional density functions are undersmoothed. And the first step requires a higher order kernel in dependence of the dimension of the regressors. For the second step, Assumption 3(v) implies that one may either use the same (higher order) kernel and bandwidth as for the first step, or alternatively a second-order kernel, requiring a smaller bandwidth $h_2 < h_1$. In the latter case, the estimation error of the first step density estimators is first-order asymptotically negligible.[7] In Assumption 3(v), $nhh_1^{2r_1} \to 0$ and $nhh_2^{2r_2} \to 0$ are the undersmoothing conditions for the limiting distribution of the estimators to be normal and centered at zero. To implement our methods in practice, an alternative set of sufficient conditions for the nonparametric tuning parameters in Assumption 3(v) is the following. Let the positive bandwidths vanish at a polynomial rate; that is, $h_1 = C_1 n^{-a}$ and $h_1 = C_1 n^{-b}$. If $h = \min\{h_1, h_2\} = h_2$, then Assumption 3(v) implies $r_1 > s$, $\max\left\{1 - 2r_1 a, \frac{1}{2r_2+1}, \frac{1-a}{2r_2}\right\} < b < 1 - 2sa$, and $\frac{1}{2r_1+1} \leq a < \min\left\{\frac{2r_2-1}{4r_2 s-1}, \frac{r_2}{s(2r_2+1)}\right\}$. For an example of $s = 3$, one may choose $r_1 = 4$. Then $r_2 = 2$, $a = 0.12$, and $b = 0.25$ satisfy the above conditions. The following theorem provides the main result of the paper, namely the asymptotic normality of our estimator.

Theorem 1. ***(Asymptotics for the nonparametric case)*** *Suppose Assumptions 1,2 and 3 hold with* $r \geq \max\{r_1, r_2\}$. *Denote by* $R(k) \equiv \int_{-\infty}^{\infty} k^2(u)du$ *and* $g(d, M_i, X_i) \equiv E[Y|D = d, M_i, X_i]$. *Then* $\sqrt{nh}\left(\widehat{\mu}(d,d) - \mu(d,d)\right) =$

[6] A kernel k is of order r if $\int k(u)du = 1$, $\int u^l k(u)du = 0$ for $0 < l < r$, and $\int |u^r k(u)|du < \infty$.

[7] Furthermore, the convergence rate is slower than the rate when we use the same higher order kernel for both steps.

$\sqrt{h/n}\sum_{i=1}^{n}\varphi^{np}_{\mu(d,d)}(Y_i,D_i,X_i;h_1,h_2)+o_p(1)\xrightarrow{d}\mathcal{N}(0,V_d)$, *where*

$$\varphi^{np}_{\mu(d,d)}(Y_i,D_i,X_i;h_1,h_2)\equiv(Y_i-\mu(d,d))\frac{K_{2,h_2}(D_i-d)}{f_D(d|X_i)}$$
$$-(E[Y|D=d,X_i]-\mu(d,d))\frac{K_{1,h_1}(D_i-d)}{f_D(d|X_i)}\quad \textit{and}$$
$$V_d\equiv\begin{cases}E\left[\mathrm{var}[Y|D=d,X]/f_D(d|X)\right]R(k_2) & \textit{if } h=h_1=h_2 \textit{ and } k_1=k_2,\\ E\left[E\left[(Y-\mu(d,d))^2|D=d,X\right]/f_D(d|X)\right]R(k_2) & \textit{if } h=h_2<h_1,\end{cases}$$

and $\sqrt{nh}\left(\hat{\mu}(d,d')-\mu(d,d')\right)=\sqrt{h/n}\sum_{i=1}^{n}\varphi^{np}_{\mu(d,d')}(Y_i,D_i,M_i,X_i;h_1,h_2)+o_p(1)\xrightarrow{d}\mathcal{N}(0,V_{dd'})$, *where*

$$\varphi^{np}_{\mu(d,d')}(Y_i,D_i,M_i,X_i;h_1,h_2)$$
$$\equiv\left(\left(Y_i-\mu(d,d')\right)K_{2,h_2}(D_i-d)\right.$$
$$\left.-\left(g(d,M_i,X_i)-\mu(d,d')\right)K_{1,h_1}(D_i-d)\right)\frac{f_D(d'|M_i,X_i)}{f_D(d|M_i,X_i)f_D(d'|X_i)}$$
$$+\left(g(d,M_i,X_i)-E\left[g(d,M,X_i)|D=d',X_i\right]\right)\frac{K_{1,h_1}(D_i-d')}{f_D(d'|X_i)}\quad \textit{and}$$

$$V_{dd'}\equiv\begin{cases}\left(E\left[\mathrm{var}[Y|D=d,X]\frac{f_D^2(d'|M,X)}{f_D(d|M,X)f_D^2(d'|X)}\right]\right.\\ \quad\left.+E\left[\mathrm{var}[g(d,M,X)|D=d',X]/f_D(d'|X)\right]\right)R(k_2), & \textit{if } h=h_1=h_2 \textit{ and } k_1=k_2,\\ E\left[E\left[\left(Y-\mu(d,d')\right)^2|D=d,X\right]\frac{f_D^2(d'|M,X)}{f_D(d|M,X)f_D^2(d'|X)}\right]R(k_2), & \textit{if } h=h_2<h_1.\end{cases}$$

Following Theorem 1, we have the following corollary regarding the asymptotics of the estimators of natural direct and indirect effects.

Corollary 1. *Suppose Assumptions 1,2, and 3 hold with* $r\geq\max\{r_1,r_2\}$. *Then*

$$\sqrt{nh}(\hat{\theta}_{d,d'}(d)-\theta_{d,d'}(d))\equiv\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\varphi^{np}_{\theta_{d,d'}(d)}(Y_i,D_i,M_i,X_i;h_1,h_2)+o_p(1),$$
$$\sqrt{nh}(\hat{\theta}_{d,d'}(d')-\theta_{d,d'}(d'))\equiv\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\varphi^{np}_{\theta_{d,d'}(d')}(Y_i,D_i,M_i,X_i;h_1,h_2)+o_p(1),$$
$$\sqrt{nh}(\hat{\delta}_{d,d'}(d)-\delta_{d,d'}(d))\equiv\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\varphi^{np}_{\delta_{d,d'}(d)}(Y_i,D_i,M_i,X_i;h_1,h_2)+o_p(1),$$
$$\sqrt{nh}(\hat{\delta}_{d,d'}(d)-\delta_{d,d'}(d))\equiv\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\varphi^{np}_{\delta_{d,d'}(d)}(Y_i,D_i,M_i,X_i;h_1,h_2)+o_p(1),$$

where

$$\varphi^{np}_{\theta_{d,d'}(d)}(Y_i,D_i,M_i,X_i;h_1,h_2)=\varphi^{np}_{\mu(d,d)}(Y_i,D_i,X_i;h_1,h_2)-\varphi^{np}_{\mu(d',d)}(Y_i,D_i,M_i,X_i;h_1,h_2),$$
$$\varphi^{np}_{\theta_{d,d'}(d')}(Y_i,D_i,M_i,X_i;h_1,h_2)=\varphi^{np}_{\mu(d,d')}(Y_i,D_i,M_i,X_i;h_1,h_2)-\varphi^{np}_{\mu(d',d')}(Y_i,D_i,X_i;h_1,h_2),$$
$$\varphi^{np}_{\delta_{d,d'}(d)}(Y_i,D_i,M_i,X_i;h_1,h_2)=\varphi^{np}_{\mu(d,d)}(Y_i,D_i,X_i;h_1,h_2)-\varphi^{np}_{\mu(d,d')}(Y_i,D_i,M_i,X_i;h_1,h_2),$$
$$\varphi^{np}_{\theta_{d,d'}(d')}(Y_i,D_i,M_i,X_i;h_1,h_2)=\varphi^{np}_{\mu(d',d)}(Y_i,D_i,M_i,X_i;h_1,h_2)-\varphi^{np}_{\mu(d',d')}(Y_i,D_i,X_i;h_1,h_2).$$

Note that the asymptotic variance of $\hat{\theta}_{d,d'}(d)$ is $V_d+V_{dd'}-2\lim_{n\to\infty}h\mathrm{cov}(\varphi^{np}_{\mu(d,d)},\varphi^{np}_{\mu(d',d)})$, whose explicit form is notationally complicated and does not provide any new insights, so we ignore it. The same argument applies to the other three estimators. Inference may be based on a sample analog estimator. For example, given uniformly consistent estimators $\hat{\mathrm{E}}[Y|D=d,X=x]$ and $\hat{\mathrm{E}}[Y|D=d,M=m,X=x]$ for $E[Y|D=d,X=x]$ and $E[Y|D=d,M=m,X=x]$, respectively, a

consistent estimator for $\mathcal{V}_{\theta_{d,d'}(d)}$ is

$$\widehat{\mathcal{V}}_{\theta_{d,d'}(d)} = \frac{h}{n}\sum_{i=1}^{n}(\widehat{\varphi}^{np}_{\mu(d,d)}(Y_i, D_i, X_i; h_1, h_2) - \widehat{\varphi}^{np}_{\mu(d',d)}(Y_i, D_i, M_i, X_i; h_1, h_2))^2,$$

where $\widehat{\varphi}^{np}_{\mu(d,d)}(Y_i, D_i, X_i; h_1, h_2)$ and $\widehat{\varphi}^{np}_{\mu(d',d)}(Y_i, D_i, M_i, X_i; h_1, h_2)$ are obtained by replacing the unknown functions or parameters in $\varphi^{np}_{\mu(d,d)}(Y_i, D_i, X_i; h_1, h_2)$ and $\varphi^{np}_{\mu(d',d)}(Y_i, D_i, M_i, X_i; h_1, h_2)$ with their uniform consistent estimators. This applies both when a single bandwidth is used such that $h = h_1 = h_2$ and $k_1 = k_2$ as well as when $h = h_1 < h_2$. Consistent variance estimators for the other point estimators of the natural direct and indirect effects can be obtained analogously.

As an alternative to basing variance estimation on the sample analogs of Theorem 1, one may apply bootstrap methods. Bootstrapping is known to be valid for local constant estimators; see Horowitz (2001). In the proof of Theorem 1, we can replace the random sample $\{(Y_i, M_i, D_i, X_i)\}_{i=1,\dots,n}$ with the bootstrap sample $\{(Y_i^*, M_i^*, D_i^*, X_i^*)\}_{i=1,\dots,n}$ and replace the population distribution p and E with the empirical distribution p^* and E^*.[8] Thus the bootstrap is valid in this context.

Our theory so far only considered the case in which all elements in X and M are continuous variables. We subsequently briefly discuss the inclusion of discrete variables. Consider a discrete covariate, $\tilde{X}$, that only takes a finite number of values and enters the conditioning set in Assumptions 1 and 2 in addition to the continuously distributed X. The conditional density of $D = d$ given the covariates may be estimated by

$$\widehat{f}_D(d|X_i, \tilde{X}_i) = \frac{\sum_{j=1}^{n} 1(\tilde{X}_j = \tilde{X}_i)K_{1,h_1}(X_j - X_i)K_{1,h_1}(D_j - d)}{\sum_{j=1}^{n} 1(\tilde{X}_j = \tilde{X}_i)K_{1,h_1}(X_j - X_j)},$$

that is, in subcells defined upon the values of $\tilde{X}$. Analogously, $\widehat{f}_D(d|M_i, X_i, \tilde{X}_i)$ is obtained. Replacing $\widehat{f}_D(d|X_i)$ and $\widehat{f}_D(d|M_i, X_i)$ in (12) by $\widehat{f}_D(d|X_i, \tilde{X}_i)$, and $\widehat{f}_D(d|M_i, X_i, \tilde{X}_i)$, respectively, allows estimating $\mu(d, d)$ and $\mu(d, d')$. When substituting $f_{DMX}(d, m, x)$ and $E[Y|D = d, M = m, X = x]$ by $f_{DMX\tilde{X}}(d, m, x, \tilde{x})$ and $E[Y|D = d, M = m, X = x, \tilde{X} = \tilde{x}]$, respectively, in Assumption 3, our previous asymptotic results remain valid.[9]

We conclude this section by considering semiparametric estimation of $\mu(d, d)$ and $\mu(d, d')$, in which the generalized propensity scores $f_D(d|X)$ and $f_D(d|M, X)$ are parametrically specified. To this end, we invoke the following assumption on the first step estimation of the generalized propensity scores.

Assumption 4. (Parametric generalized propensity scores):

(i) The estimator $\widehat{\gamma}_x$ of the generalized propensity score model $f_D(d|x; \gamma_x)$, $\gamma_x \in \Gamma_x \subseteq \mathcal{R}^{s_x}$, satisfies $\sup_{x\in\mathcal{X}}|f_D(d|x; \widehat{\gamma}_x) - f_D(d|x; \gamma_{x0})| = O_p(n^{-1/2})$, where $\gamma_{x0} \in \Gamma_x$ such that $f_D(d|x) = f_D(d|x; \gamma_{x0})$ for all $x \in \mathcal{X}$.

(ii) The estimator $\widehat{\gamma}_{mx}$ of the generalized propensity score model $f_D(d|m, x; \gamma_{mx})$, $\gamma_{mx} \in \Gamma_{mx} \subseteq \mathcal{R}^{s_{mx}}$, satisfies $\sup_{m\in\mathcal{M},x\in\mathcal{X}}|f_D(d|m, x; \widehat{\gamma}_{mx}) - f_D(d|m, x; \gamma_{mx0})| = O_p(n^{-1/2})$ where $\gamma_{mx0} \in \Gamma_{mx}$, such that $f_D(d|m, x) = f_D(d|m, x; \gamma_{mx0})$ for all $m \in \mathcal{M}$ and $x \in \mathcal{X}$.

(iii) $f_D(d|x)$ and $f_D(d|m, x)$ are uniformly bounded above and bounded away from zero on $\mathcal{D} \times \mathcal{M} \times \mathcal{X}$.

A sufficient condition for Assumption 4 is the following. Suppose that the joint density function of D, M and X, $f_{DMX}(d, m, x)$ is uniformly bounded above and bounded away from zero and follows a parametric model such that $|f_{DMX}(d, m, x) - f_{DMX}(d, m, x; \widehat{\gamma})|$ is $O_p(n^{-1/2})$ uniformly. $\widehat{\gamma}$ is a root-n consistent estimator for γ_0 (typically based on maximum likelihood) with $f_{DMX}(d, m, x) = f_{DMX}(d, m, x; \gamma_0)$. Let $f_X(x)$, $f_{DX}(d, x)$, $f_{MX}(m, x)$ be the marginal density functions. Then $f_D(d|x) = f_{DX}(d, x)/f_X(x)$ and $f_D(d|m, x) = f_{DMX}(d, m, x)/f_{MX}(m, x)$, which can be consistently estimated by $f_D(d|x; \widehat{\gamma}) = f_{DX}(d, x; \widehat{\gamma})/f_X(x; \widehat{\gamma})$ and $f_D(d|m, x; \widehat{\gamma}) = f_{DMX}(d, m, x; \widehat{\gamma})/f_{MX}(m, x; \widehat{\gamma})$. Semiparametric estimators for $\mu(d, d)$

[8] Lemma 3.1 in Powell, Stock, and Stoker (1989) and the asymptotic linear representation for the U-statistic hold for the bootstrap estimator. The Lyapounov condition holds by the same argument.

[9] Note that s_x and s_m correspond to the numbers of continuous variables in X and M, respectively—that is, without the discrete covariate $\tilde{X}$.

and $\mu(d,d')$ are given by

$$\begin{aligned}\widehat{\mu}(d,d) &= \sum_{i=1}^{n} \frac{Y_i K_{2,h_2}(D_i-d)}{\widehat{f}_D(d|X_i;\widehat{\gamma}_x)} \Bigg/ \sum_{i=1}^{n} \frac{K_{2,h_2}(D_i-d)}{\widehat{f}_D(d|X_i;\widehat{\gamma}_x))},\\ \widehat{\mu}(d,d') &= \sum_{i=1}^{n} \frac{Y_i K_{2,h_2}(D_i-d)}{\widehat{f}_D(d|M_i,X_i;\widehat{\gamma}_{mx}))} \cdot \frac{\widehat{f}_D(d'|M_i,X_i;\widehat{\gamma}_{mx})}{\widehat{f}_D(d'|X_i;\widehat{\gamma}_x)} \Bigg/ \sum_{i=1}^{n} \frac{K_{2,h_2}(D_i-d)}{\widehat{f}_D(d|M_i,X_i;\widehat{\gamma}_{mx})} \cdot \frac{\widehat{f}_D(d'|M_i,X_i;\widehat{\gamma}_{mx})}{\widehat{f}_D(d'|X_i;\widehat{\gamma}_x)}.\end{aligned} \tag{13}$$

By invoking Assumption 4, the asymptotic theory for these estimators simplifies considerably when compared to the nonparametric case; see Theorem 2 below.

Theorem 2 (Asymptotics for the semiparametric case)**.** *Suppose Assumptions 1-3(i)–(iv), and 4 hold with $r \geq r_2$. Let the order of the kernel $r_2 = 2$. The bandwidth h_2 satisfy $h_2 \to 0$, $nh_2 \to \infty$, and $nh_2^5 \to 0$. Then*

$$\begin{aligned}&\sqrt{nh_2}\left(\widehat{\mu}(d,d) - \mu(d,d)\right)\\ &= \sqrt{\frac{h_2}{n}} \sum_{i=1}^{n} (Y_i - \mu(d,d)) \frac{K_{2,h_2}(D_i-d)}{f_D(d|X_i)} + o_p(1) \xrightarrow{d} \mathcal{N}(0, V_d),\end{aligned}$$

where $V_d = E\left[E\left[(Y-\mu(d,d))^2|D=d,X\right]/f_D(d|X)\right]R(k_2)$ *and*

$$\begin{aligned}&\sqrt{nh_2}\left(\widehat{\mu}(d,d') - \mu(d,d')\right)\\ &= \sqrt{\frac{h_2}{n}} \sum_{i=1}^{n} \left(Y_i - \mu(d,d')\right) \frac{K_{2,h_2}(D_i-d) f_D(d'|M_i,X_i)}{f_D(d|M_i,X_i) f_D(d'|X_i)} + o_p(1) \xrightarrow{d} \mathcal{N}(0, V_{dd'}),\end{aligned}$$

where $V_{dd'} = E\left[E\left[\left(Y-\mu(d,d')\right)^2|D=d,M,X\right]\frac{f_D^2(d'|M,X)}{f_D(d|M,X)f_D^2(d'|X)}\right]R(k_2)$.

The condition $nh_2^5 \to 0$ in Theorem 2 is the undersmoothing condition for the semiparametric estimators. A corollary similar to Corollary 1 for the asymptotics for semiparametric estimators for natural direct and indirect estimators can be obtained similarly, so we omit the details. The main advantage of the semiparametric approach over the fully nonparametric estimator is that it circumvents the curse of dimensionality problem when the dimensions of X and/or M are large. On the downside, misspecifications of the generalized propensity scores generally result in inconsistent estimators of potential outcomes and effects.

5 | SIMULATION STUDY

This section provides a simulation study to investigate the finite sample behavior of our semi- and nonparametric methods based on the following data generating process:

$$\begin{aligned}&Y = 0.3D + 0.3\,M + \alpha DM + 0.3X + \beta D^3 + U,\\ &M = 0.3D + 0.3X + V, \quad D = 0.3X + W,\\ &X \sim \text{uniform}(-1.5, 1.5),\ U, V, W \sim \text{uniform}(-2, 2), \text{independently of each other.}\end{aligned}$$

Outcome Y is a function of the observed variables D, M, X and an unobserved term U. α gauges the interaction effect between D and M. $\alpha = 0$ satisfies the assumption of no interaction as discussed in Robins (2003), implying that the direct effect $\theta_{d,d'}(d) = \theta_{d,d'}(d')$ in Equation 2 and the indirect effect $\delta_{d,d'}(d) = \delta_{d,d'}(d')$ in Equation 3. In contrast, for $\alpha \neq 0$, direct and indirect effects are heterogeneous. β determines whether the direct effect of D on Y is linear (β=0) or nonlinear, namely cubic ($\beta \neq 0$). Mediator M is a function of D, X and the unobservable V. Note that the indirect effect is linear, as M is linear in D and Y is linear in M. Treatment D is linearly determined by X and the unobservable W. The covariate X, which confounds the treatment–outcome, treatment–mediator, and mediator–outcome relations, is continuously uniformly distributed with support ranging from −1.5 to 1.5. Finally, the unobservables follow uniform distributions with support ranging from −2 to 2. They are statistically independent of each other as well as of X. In our

TABLE 1 Simulations $\alpha = 0.5$, $\beta = 0$

	$\widehat{\theta}_{d,0}(d)$			$\widehat{\theta}_{d,0}(0)$			$\widehat{\delta}_{d,0}(d)$			$\widehat{\delta}_{d,0}(0)$		
	abias	**SD**	**RMSE**	**abias**	**SD**	**RMSE**	**abias**	**SD**	**RMSE**	**abias**	**SD**	**RMSE**
$n = 1,000$												
OLS	0.124	0.035	0.130	0.000	0.035	0.035	0.124	0.013	0.125	0.001	0.013	0.013
W np	0.020	0.057	0.062	0.062	0.056	0.086	0.077	0.010	0.077	0.039	0.007	0.040
W np us	0.016	0.101	0.103	0.044	0.100	0.113	0.048	0.035	0.060	0.023	0.024	0.034
W p	0.059	0.059	0.086	0.058	0.058	0.083	0.011	0.020	0.024	0.006	0.015	0.016
W p us	0.050	0.106	0.118	0.049	0.105	0.117	0.003	0.024	0.024	0.002	0.019	0.019
$n = 4,000$												
OLS	0.124	0.017	0.126	0.000	0.017	0.017	0.124	0.006	0.124	0.000	0.006	0.006
W np	0.016	0.038	0.044	0.054	0.037	0.069	0.065	0.008	0.065	0.034	0.005	0.034
W np us	0.021	0.063	0.067	0.043	0.062	0.079	0.048	0.021	0.052	0.026	0.014	0.029
W p	0.050	0.039	0.065	0.050	0.038	0.064	0.005	0.011	0.013	0.001	0.008	0.008
W p us	0.049	0.065	0.084	0.049	0.064	0.083	0.003	0.014	0.014	0.001	0.011	0.011

Note. "abias," "SD," and "RMSE" report the the average absolute bias, standard deviation, and root mean squared error, respectively, of the effects across all treatment values $d \in \{-1.5, -1.4, \ldots 1.4, 1.5\}$ and $d' = 0$. "OLS," "W np," "W np us," "W p," and "W p us" refer to linear regression, nonparametric weighting, nonparametric weighting with undersmoothing in the kernel procedures, weighting with a parametric generalized propensity score, and weighting with a parametric generalized propensity score and undersmoothing in the kernel function, respectively.

simulation design, the ATE corresponds to $\Delta_{d,d'} = 0.39(d - d') + 0.3\alpha(d^2 - d'^2) + \beta(d^3 - d'^3)$. The direct effects are given by $\theta_{d,d'}(d) = 0.3(d - d') + 0.3\alpha(d^2 - d'^2) + \beta(d^3 - d'^3)$ and $\theta_{d,d'}(d') = 0.3(d - d') + \beta(d^3 - d'^3)$, and the indirect effects by $\delta_{d,d'}(d) = 0.09(d - d') + 0.3\alpha(d^2 - d'^2)$ and $\delta_{d,d'}(d') = 0.09(d - d')$.

We consider 1,000 simulations and two sample sizes $n = 1,000, 4,000$ to investigate the performance of our nonparametric weighting approach based on Equation (12). As the dimension of (D, X, M) is equal to $s = 3$ (see Section 4) in our simulation, we set the orders of the Epanechnikov kernels in Equations 11 and 12 to $r_1 = 4$ and $r_2 = 2$, respectively. Furthermore, the bandwidth h_1 is determined by multiplying the respective standard deviations of D, X, M by $C_1 n^{-0.12}$, where $C_1 = 3.03$ is the constant term in a Silverman (1986)-type rule of thumb for fourth-order Epanechnikov kernels. Analogously, h_2 is obtained using $C_2 n^{-0.25}$, with $C_2 = 2.34$ being the constant for second-order Epanechnikov kernels. We note that these choices of r_1, r_2, h_1, h_2 satisfy the regularity conditions in Assumption 3 required for the satisfaction of Theorem 1.

Furthermore, we consider semiparametric weighting based on parametric estimation of the generalized propensity scores in Equation 13. To this end we (incorrectly) assume D to be normally distributed given X or given (X, M), respectively. Bandwidth h_2 corresponds to $C_2 n^{-0.25}$, with $C_2 = 2.34$. For all kernel-based computations, we use the "np" package by Hayfield and Racine (2008) for the statistical software "R." Besides estimation using bandwidths based on the rule of thumb, we consider undersmoothed versions, in which bandwidths of all kernel procedures are divided by 2. For comparison, in addition we estimate the direct and indirect effects based on linear ordinary least squares (OLS) regressions of the mediator on a constant, the treatment, and covariate and of the outcome on a constant, the treatment, the mediator, and the covariate, respectively. Concerning the definition of the direct and indirect effects, we set $d' = 0$. For d, we consider a sequence of values defined by an equidistant grid between (and including) -1.5 and 1.5 with step size 0.1 (i.e., $d \in \{-1.5, -1.4, \ldots 1.4, 1.5\}$; however, without including 0 for obvious reasons.

Table 1 reports the averages of the absolute bias (abias), standard deviation (SD), and root mean squared error (RMSE) for each effect under $\alpha = 0.5$ (effect heterogeneity) and $\beta = 0$ (fully linear model), where averaging is over all treatment comparisons $(d - d')$ considered. Not surprisingly, the OLS-based estimators (OLS) have the lowest standard deviations of all methods due to their parametric assumptions. On the downside, the OLS estimates of $\theta(d)$ and $\delta(d)$ are nonnegligibly biased under either sample size due to the omission of the treatment–mediator interactions. In contrast, the nonparametric weighting estimator with rule-of-thumb bandwidths (W np) is considerably less biased. Undersmoothing (W np us) generally entails an even lower absolute bias but, as expected, a higher standard deviation. A qualitatively similar pattern is observed for semiparametric weighting with a parametric first step (W p). Undersmoothing (W p us), which in the semiparametric case only concerns h_2, reduces the absolute bias and increases the standard deviation. We also note that the semi- and nonparametric versions do not uniformly dominate each other in terms of RMSE across the effects and sample sizes considered.

TABLE 2 Simulations $\alpha = 0$, $\beta = 0.25$

	$\widehat{\theta}_{d,0}(d)$			$\widehat{\theta}_{d,0}(0)$			$\widehat{\delta}_{d,0}(d)$			$\widehat{\delta}_{d,0}(0)$		
	abias	**SD**	**RMSE**	**abias**	**SD**	**RMSE**	**abias**	**SD**	**RMSE**	**abias**	**SD**	**RMSE**
$n = 1,000$												
OLS	0.280	0.029	0.282	0.280	0.029	0.282	0.001	0.011	0.011	0.001	0.011	0.011
W np	0.099	0.055	0.117	0.097	0.055	0.115	0.035	0.009	0.036	0.038	0.008	0.039
W np us	0.043	0.096	0.106	0.041	0.097	0.105	0.021	0.025	0.033	0.023	0.024	0.034
W p	0.064	0.057	0.090	0.066	0.058	0.091	0.015	0.018	0.023	0.007	0.015	0.016
W p us	0.024	0.101	0.105	0.026	0.101	0.106	0.004	0.020	0.021	0.002	0.018	0.019
$n = 4,000$												
OLS	0.281	0.015	0.281	0.281	0.015	0.281	0.000	0.006	0.006	0.000	0.006	0.006
W np	0.064	0.036	0.074	0.061	0.036	0.072	0.031	0.006	0.031	0.034	0.005	0.034
W np us	0.035	0.059	0.069	0.033	0.059	0.068	0.024	0.014	0.028	0.026	0.014	0.029
W p	0.023	0.037	0.046	0.025	0.037	0.048	0.007	0.009	0.012	0.001	0.008	0.008
W p us	0.034	0.062	0.072	0.036	0.062	0.073	0.001	0.011	0.011	0.001	0.011	0.011

Note. "abias," "SD," and "RMSE" report the the average absolute bias, standard deviation, and root mean squared error, respectively, of the effects across all treatment values $d \in \{-1.5, -1.4, \ldots 1.4, 1.5\}$ and $d' = 0$. "OLS," "W np," "W np us," "W p," and "W p us" refer to linear regression, nonparametric weighting, nonparametric weighting with undersmoothing in the kernel procedures, weighting with a parametric generalized propensity score, and weighting with a parametric generalized propensity score and undersmoothing in the kernel function, respectively.

Table 2 gives the average statistics over all treatment comparisons $(d - d')$ for $\alpha = 0$ (effect homogeneity) and $\beta = 0.25$ (nonlinear direct effects). The OLS estimates of the direct effects are severely biased due to the cubic effect of D in the outcome model, whereas the indirect effect estimates are unbiased, as they are indeed linear. In contrast, the absolute biases of both the semi- and nonparametric weighting estimators for the direct effects are considerably smaller and decreasing in the sample size. Again, undersmoothing in many cases entails a lower absolute bias than relying on rule-of-thumb bandwidths, but leads to higher standard deviations. Interestingly, the semiparametric versions (W p, W p us) are quite competitive both in terms of small absolute biases and RMSEs, despite incorrectly assuming normality. Apparently, the misspecification of the generalized propensity score does not entail important biases as long as bandwidth h_2 is sufficiently small.

Finally, Table 3 provides the results when setting $\alpha = 0.5, \beta = 0.25$ (effect heterogeneity and nonlinear direct effects). Three out of four OLS effect estimates exhibit important biases, while both the semi- and nonparametric weighting estimators are less biased and superior to OLS in terms of average RMSEs under either sample size. All in all, the simulations demonstrate the merits of our methods in terms of robustness to deviations from specific parametric assumptions. This, however, comes at an efficiency cost which decreases in the sample size. The results suggest that our methods perform decently in sample sizes with several thousand observations (or more), which is quite common in empirical research.

6 | EMPIRICAL ILLUSTRATION

We apply our method to the Job Corps study, which was conducted in the mid-1990s to assess the publicly funded US Job Corps program and used an experimental design in which access to Job Corps was assigned at random. The Job Corps program targets individuals who are between 16 and 24 years old, legally reside in the USA, and come from low-income households. Participants received approximately 1,200 hours of vocational training and education, housing, and board over an average duration of eight months. Schochet, Burghardt, and Glazerman (2001) and Schochet, Burghardt, and McConnell (2008) discuss in detail the study design and report the average effects of program assignment on a broad range of outcomes. Their findings suggest that Job Corps increases educational attainment, reduces criminal activity, and increases employment and earnings, at least for some years after the program.

Several previous studies investigated various causal mechanisms of the Job Corps program and found significant direct or indirect effects, depending on the mediator and outcome variables considered. Flores and Flores-Lagunes (2009) find a positive direct effect of program assignment on earnings when controlling for the mediator work experience which they assume to be conditionally exogenous given observed covariates. Also, Huber (2014) invokes a selection on observables assumption and estimates a positive direct health effect when controlling for the mediator employment. Frölich and Huber (2017) use an IV strategy based on two instruments to disentangle the earnings effect of being enrolled in Job Corps into an indirect effect via hours worked and a direct effect (likely related to a change in human capital). The results point to

TABLE 3 Simulations $\alpha = 0.5$, $\beta = 0.25$

	$\widehat{\theta}_{d,0}(d)$			$\widehat{\theta}_{d,0}(0)$			$\widehat{\delta}_{d,0}(d)$			$\widehat{\delta}_{d,0}(0)$		
	abias	**SD**	**RMSE**	**abias**	**SD**	**RMSE**	**abias**	**SD**	**RMSE**	**abias**	**SD**	**RMSE**
$n = 1,000$												
OLS	0.298	0.037	0.303	0.280	0.037	0.283	0.124	0.013	0.125	0.001	0.013	0.013
W np	0.100	0.061	0.122	0.114	0.060	0.132	0.076	0.011	0.077	0.038	0.008	0.039
W np us	0.044	0.102	0.112	0.056	0.101	0.120	0.047	0.035	0.060	0.023	0.024	0.034
W p	0.068	0.063	0.097	0.067	0.062	0.095	0.016	0.021	0.029	0.007	0.015	0.017
W p us	0.025	0.107	0.112	0.026	0.106	0.111	0.004	0.024	0.025	0.002	0.019	0.019
$n = 4,000$												
OLS	0.299	0.018	0.300	0.281	0.018	0.282	0.124	0.007	0.124	0.000	0.007	0.007
W np	0.064	0.039	0.076	0.078	0.038	0.089	0.065	0.008	0.065	0.034	0.005	0.034
W np us	0.035	0.063	0.073	0.049	0.062	0.083	0.047	0.021	0.052	0.026	0.014	0.029
W p	0.029	0.040	0.053	0.029	0.040	0.053	0.008	0.012	0.015	0.002	0.008	0.008
W p us	0.035	0.065	0.077	0.036	0.065	0.077	0.003	0.014	0.014	0.001	0.011	0.011

Note. "abias," "SD," and "RMSE" report the the average absolute bias, standard deviation, and root mean squared error, respectively, of the effects across all treatment values $d \in \{-1.5, -1.4, \ldots 1.4, 1.5\}$ and $d' = 0$. "OLS," "W np," "W np us," 'W p," and "W p us" refer to linear regression, nonparametric weighting, nonparametric weighting with undersmoothing in the kernel procedures, weighting with a parametric generalized propensity score, and weighting with a parametric generalized propensity score and undersmoothing in the kernel function, respectively.

the existence of an indirect rather than a direct mechanism. Using a partial identification approach allowing for mediator endogeneity, Flores and Flores-Lagunes (2010) derive bounds for direct and indirect effects of Job Corps assignment on employment and earnings mediated by the achievement of a GED, high school degree, or vocational degree. Under their strongest set of bounding assumptions, the results suggest a positive effect on labor market outcomes even net of the indirect mechanism via obtaining a degree.

While these previous contributions consider binary treatment definitions, our interest lies in the effect of different doses of participation in Job Corps on an outcome variable capturing criminal behavior, namely the number of arrests. Our continuous treatment definition follows Flores et al. (2012), who assess the total effect of length of exposure to academic and vocational instruction on earnings. In contrast, our mediation analysis investigates whether the time spent in Job Corps affects the number of arrests indirectly through employment or "directly"—that is, through any other causal mechanisms. More precisely, our treatment variable D is defined as the total hours spent either in academic or vocational classes in the 12 months following the program assignment according to the survey. While access to the program was randomly assigned, the decision to actually take the treatment was endogenous and thus prone to selection, both at the extensive margin (whether to join Job Corps or not) and at the intensive margin (how many hours to consume). The mediator M is the proportion of weeks employed in the second year, while the outcome variable Y corresponds to the number of times the individual was arrested by the police in the fourth year after the random assignment.

Schochet et al. (2008) report that Job Corps significantly reduced arrest and conviction rates, as well as time spent incarcerated. Our approach adds to these findings in two dimensions. First, we document that the effect on the number of arrests is highly nonlinear in the treatment dose, with significant reductions in arrests only materializing after a non-negligible amount of hours in Job Corps. A binary treatment definition would not permit discovering this nonlinearity. Second, our mediation analysis disentangles the total reduction into an indirect component due to Job Corps-induced employment and a (direct) remainder effect of the program, which allows assessing the relative importance of different causal mechanisms.

For identification, we invoke sequential conditional independence of the treatment and the mediator as outlined in Section 3 based on a rich set of pretreatment covariates X, which overlaps with the control variables of Flores et al. (2012).[10] Specifically, we control for individual characteristics like age, gender, ethnicity, language competency, education, marital status, household size and income, previous receipt of social aid, and family background (e.g., parents' education), as well as health and health-related behavior at baseline. Conditioning on such a rich set of socioeconomic variables appears important, as identification relies on successfully controlling for all confounders jointly influencing at least two out of the three variables time in treatment, employment in the second year, and arrests in the fourth year. Furthermore,

[10]A control variable in Flores et al. (2012) we do not have access to is the local unemployment rate. The latter was constructed by matching county-level unemployment rates to individual postal codes of residence, which are only available in a restricted-use data set.

TABLE 4 Descriptives

Variable	Type	Mean	SD	Min	Max	Nonmissing
female	dummmy (1 if yes, 0 if no)	0.44	0.50	0.00	1.00	4,000
age	numeric	18.33	2.14	16.00	24.00	4,000
white	dummy (1 if yes, 0 if no)	0.25	0.43	0.00	1.00	4,000
black	dummy (1 if yes, 0 if no)	0.50	0.50	0.00	1.00	4,000
Hispanic	dummy (1 if yes, 0 if no)	0.17	0.38	0.00	1.00	4,000
years of education	numeric	10.05	1.54	0.00	20.00	3,945
GED diploma	dummy (1 if yes, 0 if no)	0.04	0.20	0.00	1.00	3,982
high school diploma	dummy (1 if yes, 0 if no)	0.18	0.39	0.00	1.00	3,982
native English	dummy (1 if yes, 0 if no)	0.86	0.35	0.00	1.00	3,950
divorced	dummy (1 if yes, 0 if no)	0.01	0.09	0.00	1.00	3,953
separated	dummy (1 if yes, 0 if no)	0.01	0.11	0.00	1.00	3,953
cohabiting	dummy (1 if yes, 0 if no)	0.03	0.18	0.00	1.00	3,953
married	dummy (1 if yes, 0 if no)	0.02	0.13	0.00	1.00	3,953
has children	dummy (1 if yes, 0 if no)	0.18	0.38	0.00	1.00	3,981
ever worked	dummy (1 if yes, 0 if no)	0.41	0.49	0.00	1.00	1,405
average weekly gross earnings (in USD)	numeric	19.41	98.66	0.00	2,000.00	3,999
is household head	dummy (1 if yes, 0 if no)	0.11	0.31	0.00	1.00	3,933
household size (number of people)	numeric	3.52	2.01	0.00	15.00	3,944
designated for nonresidential slot	dummy (1 if yes, 0 if no)	0.17	0.38	0.00	1.00	4,000
total household gross income	categorical (cf. Table A1)	3.51	2.21	1.00	7.00	2,508
total personal gross income	categorical (cf. Table A1)	1.11	0.48	1.00	7.00	1,774
mum's years of education	numeric	11.50	2.60	0.00	20.00	3,263
dad's years of education	numeric	11.45	2.90	0.00	20.00	2,506
dad did not work when 14	dummy (1 if yes, 0 if no)	0.06	0.23	0.00	1.00	3,575
received AFDC every month	dummy (1 if yes, 0 if no)	0.80	0.40	0.00	1.00	1,148
received public assistance every month	dummy (1 if yes, 0 if no)	0.85	0.36	0.00	1.00	946
received food stamps	dummy (1 if yes, 0 if no)	0.45	0.50	0.00	1.00	3,836
welfare receipt during childhood	categorical (cf. Table A1)	2.07	1.19	1.00	4.00	3,726
poor/fair general health status	dummy (1 if yes, 0 if no)	0.13	0.33	0.00	1.00	3,953
physical/emotional problems	dummy (1 if yes, 0 if no)	0.04	0.20	0.00	1.00	3,950
extent of marijuana use	categorical (cf. Table A1)	2.54	1.55	0.00	4.00	1,469
extent of hallucinogen use	categorical (cf. Table A1)	2.76	1.73	0.00	4.00	204
ever used other illegal drugs	dummy (1 if yes, 0 if no)	0.01	0.08	0.00	1.00	2,628
extent of smoking	categorical (cf. Table A1)	1.53	0.98	0.00	4.00	2,084
extent of alcohol consumption	categorical (cf. Table A1)	3.14	1.21	0.00	4.00	2,306
ever arrested	dummy (1 if yes, 0 if no)	0.24	0.43	0.00	1.00	3,951
times in prison	numeric	0.07	0.35	0.00	5.00	3,951
time spent by recruiter speaking of Job Corps	categorical (cf. Table A1)	2.05	0.94	1.00	4.00	3,922
extent of recruiter support	categorical (cf. Table A1)	1.59	1.07	1.00	5.00	3,911
idea about wished training	dummy (1 if yes, 0 if no)	0.85	0.35	0.00	1.00	3,944
expected hourly wage after Job Corps	numeric	9.95	6.57	5.00	96.00	1,799
expected improvement in maths	categorical (cf. Table A1)	1.32	0.53	1.00	3.00	3,916
expected improvement in reading skills	categorical (cf. Table A1)	1.53	0.65	1.00	3.00	3,932
expected improvement in social skills	categorical (cf. Table A1)	1.48	0.68	1.00	3.00	3,932
expected to be training for a job	categorical (cf. Table A1)	1.04	0.23	1.00	3.00	3,922
worried about Job Corps	dummy (1 if yes, 0 if no)	0.37	0.48	0.00	1.00	3,944
1st contact with recruiter by phone	dummy (1 if yes, 0 if no)	0.41	0.49	0.00	1.00	3,953
1st contact with recruiter in office	dummy (1 if yes, 0 if no)	0.39	0.49	0.00	1.00	2,315
expected stay in Job Corps	numeric (in months)	6.64	9.81	0.00	36.00	4,000
total hours spent in 1st year classes (D)	numeric (treatment var.)	1,194.15	964.89	0.86	5,142.86	4,000
Share of weeks employed in 2nd year (M)	numeric (in percent, mediator var.)	44.05	37.84	0.00	100.00	4,000
Number of arrests in year 4 (Y)	numeric (outcome var.)	0.15	0.62	0.00	8.00	4,000

we condition on variables that are predictive for the duration in the program, namely expectations about Job Corps and interaction with the recruiters. Such factors appear important as they are likely correlated with personality traits like motivation, which may also affect the mediator and the outcome. Finally, we include pretreatment outcome and mediator

variables that reflect labor market and criminal behavior prior to Job Corps. This permits controlling for unobserved confounders that are time constant in the sense that they only affect the mediator and the outcome through their respective pretreatment values.

We, however, acknowledge that our framework does not allow for dynamic confounding, implying that the length of treatment and/or the share of employment are affected by confounders that are themselves influenced by the initial decision to participate in the treatment at all. This would, for instance, be the case if initial treatment participation affected motivation, which in turn influenced treatment duration, employment, and criminal behavior. Even though we hope that the limited time horizon considered for the treatment (first year) and the mediator (second year) mitigates issues related to dynamic confounding, this threat to identification needs to be borne in mind when interpreting the results.

The original Job Corps data set consists of 15,386 individuals prior to program assignment, but a substantial share never enrolled in the program and dropped out of the study. We therefore consider the 10,775 observations for which both the posttreatment variables M and Y are observed in the follow-up surveys after 2 and 4 years, respectively.[11] Among these, there are cases of item nonresponse in various elements of X measured at the baseline survey, for which we account by the inclusion of missing dummies. Furthermore, and similar to Flores et al. (2012), we restrict our evaluation sample to observations with a positive treatment intensity; that is, $D > 0$, ultimately consisting of 4,000 individuals.[12] The results presented further below therefore refer to the group of treated individuals with nonmissing posttreated variables and only carry over to other groups (like the total population) if direct and indirect effects are homogeneous across individual characteristics.

Table 4 provides descriptive statistics for the pretreatment covariates as well as the treatment, mediator, and outcome variables in our evaluation sample, along with the numbers of nonmissing observations. Individuals in our evaluation sample were on average 18.33 years old at baseline when applying for Job Corps and women made up 44%. Half of the applicants were black, while whites and Hispanics accounted for 25% and 17%, respectively. Regarding education, 18% of those with nonmissing values held a high school diploma and 4% a General Education Diploma (GED). A large share of respondents (had) received public assistance or welfare benefits, pointing to economic hardship. 24% had been arrested at least once prior to program assignment (excluding minor motor vehicles violations). Concerning treatment intensity (D), individuals spent on average 1,194 hours either in academic or vocational classes in the first year after assignment. This corresponds to roughly 149 days of 8 hours. Thus individuals with a positive treatment intensity were on average almost 30 working weeks in Job Corps in the first year. The treatment distribution is right skewed as the median is somewhat lower, amounting to 966 hours in classes. Concerning the share of weeks employed in the second year (M), the individuals were on average 44.05% in employment. Finally, the average number of arrests in the fourth year (Y) amounts to 0.15. Most individuals were never arrested, while 9% were arrested at least once.

We evaluate the direct and indirect effects for 20 different values of positive treatment intensity between 100 and 2,000 hours in steps of 100 versus a rather small intensity of just 40 hours. That is, we estimate $\widehat{\theta}_{d,d'}(d)$, $\widehat{\theta}_{d,d'}(d')$, $\widehat{\delta}_{d,d'}(d)$, and $\widehat{\delta}_{d,d'}(d')$ for each of $d \in \{100, 200, \ldots, 1,900, 2,000\}$ and $d' = 40$. We therefore investigate among treated individuals whether the length of classroom education actually matters for the number of arrests relative to a minor exposure (40 hours) that corresponds to roughly one working week spent in class. This permits learning whether the treatment affects criminal behavior importantly at the extensive margin in order to judge the benefits of a more lengthy (and costly) exposure to classroom education when compared to a minimal intervention.[13] Figure 1 reports the distribution of D in our evaluation sample by means of a histogram. Due to large number of covariates, the generalized propensity scores are estimated parametrically. We therefore assume that D is conditionally log-normally distributed given X or (X, M), as it is common for nonnegative treatments; see, for instance, Imai and van Dyk (2004). As for semiparametric weighting in Section 5, estimation relies on Equation 13 and the rule of thumb for determining bandwidth h_2. We note that the obtained results are quite similar when assuming a conditional normal distribution of D (instead of log-normality) and/or applying undersmoothing by taking half of the rule-of-thumb bandwidth h_2. Inference is based on bootstrap standard errors obtained by bootstrapping the effects 999 times.

[11] Our analysis does not make use of the sample weights provided in the Job Corps data to account for the fact that, due to stratified sampling, specific groups are over- or underrepresented in the data relative to the original study population of interest.

[12] All in all, there are 5,279 observations with $D > 0$, out of which 1,279 have missing values in M and/or Y. Investigating the selectivity of missingness w.r.t. the treatment by regressing a dummy for the missingness of Y or M (or both) on D using a probit model yields a p-value of 16%.

[13] In a robustness check, we set $d' = 0$ (no classes at all) and also include observations with zero treatment intensity in our analysis. Figure A1 in Appendix A.3 displays the direct and indirect effects. The point estimates and conclusions to be drawn are similar to those presented in this section, despite the fact that the effects are defined relative to a zero treatment rather than a minor, positive treatment.

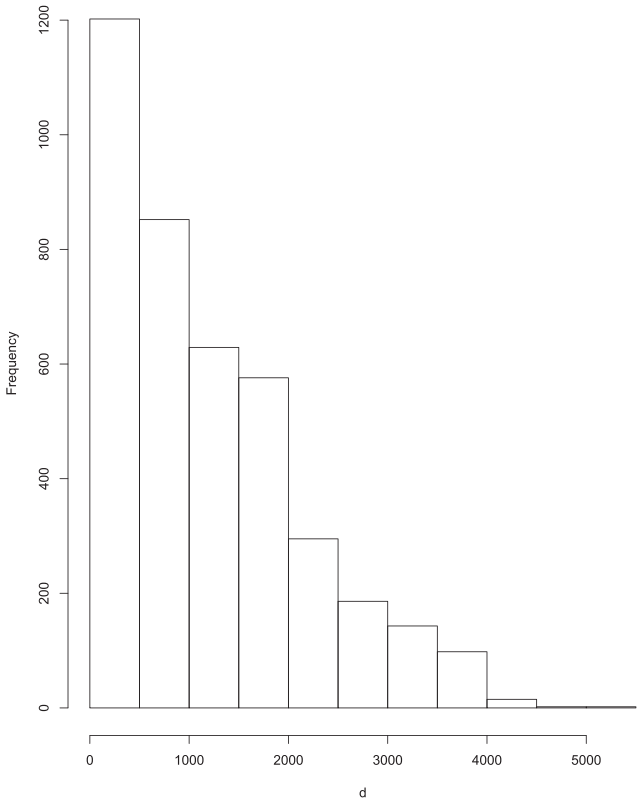

FIGURE 1 Histogram of D

To verify whether our estimates of the generalized propensity score $f_D(d|M,X)$ successfully balance the distributions of the covariates and the mediator across treatment intensities, we conduct a test that is in the spirit of Smith and Todd (2005). Specifically, we linearly regress each of the 65 elements in X (that also include missing dummies) as well as M on the log-treatment intensity, the generalized propensity score (given X and M) estimated at the sample values of D, and the score's square.[14] If (X,M) and D are not associated given the estimated propensity score such that the latter satisfies the balancing property, then the coefficient on the log-treatment should be statistically insignificant in most cases. The p-values of the coefficient averages 56.7% and is only in four regressions (6%) smaller than 5%, such that we do not find evidence for a violation of the balancing property.[15]

Furthermore, we check for common support across the generalized propensity scores estimated at the different treatment values considered in our application. Common support implies that no observation obtains too large a weight in the computation of any mean potential outcome based on weighting expressions (Equation 12), due to dividing by estimated generalized propensity scores that are close to zero. A large weight would entail a large influence of a single observation w.r.t. the estimation of some mean potential outcome, thus implying a large variance of the estimator of that mean potential outcome and ultimately of the effect of interest. See Huber, Lechner, and Wunsch (2013) for an analogous argument in the context of binary treatment evaluation. We therefore investigate the relative weights in our sample when estimating the mean potential outcomes, corresponding, for example, to $\frac{K_{2,h_2}(D_i-d)}{\hat{f}_D(d|X_i)} / \sum_{i=1}^{n} \frac{K_{2,h_2}(D_i-d)}{\hat{f}_D(d|X_i)}$ when considering $\hat{\mu}(d,d)$. For any observation, any mean potential outcome, and any $d \in \{40, 100, 200, \ldots, 1,900, 2,000\}$, the relative weight is below 1%. We therefore do not find evidence for a lack in common support.

The upper panel of Figure 2 displays the direct effects under treatment ($\hat{\theta}_{d,40}(d)$) on the left and nontreatment ($\hat{\theta}_{d,40}(40)$) on the right, which are quite heterogeneous over the range of values d. While small treatment intensities do not appear to directly reduce the number of arrests, direct effects are statistically significantly negative at the 5% level from 1,100 hours on, when the pointwise 95% confidence intervals (dashed lines) do not include zero. The effect peaks in absolute terms around 1,700 hours, reducing the number of arrests by 0.09. In relative terms, this effect is substantial, given that the average number of arrests in the fourth year is 0.15; see Table 4. The lower panel of Figure 2 provides the indirect effects under treatment ($\hat{\delta}_{d,40}(d)$) on the left and nontreatment ($\hat{\delta}_{d,40}(40)$) on the right, operating through employment.

[14]Using cubic or quartic polynomials of the propensity score yields similar results.
[15]The four variables for which balance is rejected at the 5% level are black, Hispanic, native English, and expected stay in Job Corps.

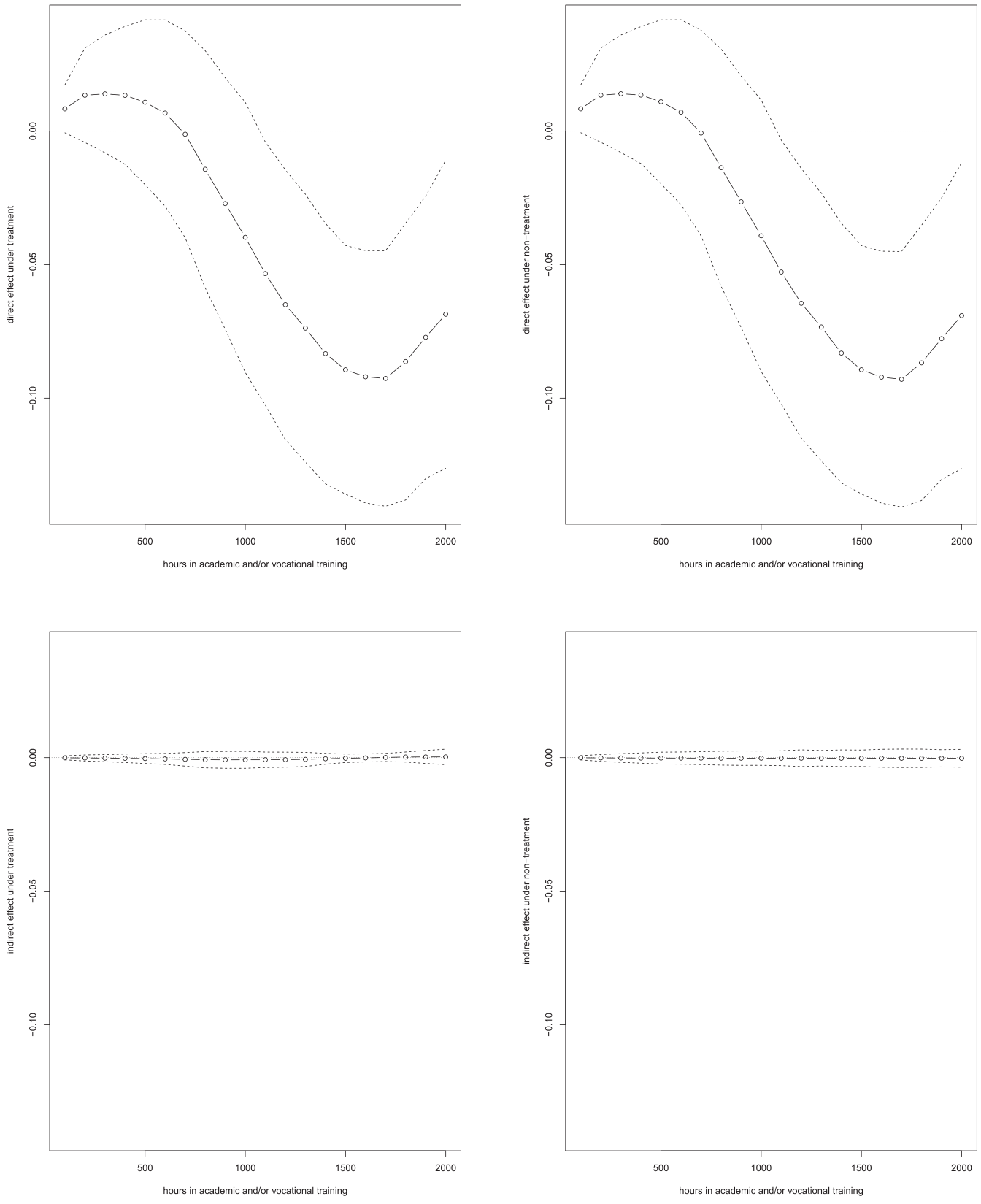

FIGURE 2 Direct effects $\widehat{\theta}_{d,40}(d)$ (top left) and $\widehat{\theta}_{d,40}(40)$ (top right) as well as indirect effects $\widehat{\delta}_{d,40}(d)$ (bottom left) and $\widehat{\delta}_{d,40}(40)$ (bottom right) for $d \in \{100, 200, \ldots, 1,900, 2,000\}$

All indirect effects are very small in absolute terms and never statistically different from zero at the 5% level. Summing up, our results point to an important direct, nonlinear reduction in the number of arrests in the fourth year as a consequence of Job Corp under a sufficiently large treatment intensity of roughly 1,100 hours or more. In contrast, the effects of program-induced employment changes on arrests are close to zero for the investigated range of treatment intensities.

To check the robustness of our findings across different methods, we also compute the direct and indirect effects for $d \in \{100, 200, \ldots, 1,900, 2,000\}$ and $d' = 40$ using the "mediate" command in the "mediation" package for "R" by Tingley et al. (2014). The latter applies regression to estimate the mediator and outcome models and simulates potential mediators and outcomes according to these models; see Imai, Keele, and Tingley (2010) for details concerning the algorithm. Among other specifications, the command permits for generalized additive models (GAM) such that the continuous regressors (be it X, M, or D) in the outcome and mediator equations are flexibly modeled by polynomial functions. We use the GAM approach and also include a polynomial of the interaction between D and M in the outcome equation to allow for heterogeneous direct and indirect effects. Appendix A.0.4 reports the effect estimates along with 95% confidence intervals based on bootstrapping 999 times. The point estimates are in line with those of our weighting estimators reported in Figure 2; however, precision is considerably lower. Finally, we redefine the outcome variable Y to be a dummy variable indicating any arrests in the fourth year ($Y = 1$) versus no arrests ($Y = 0$) and apply our semiparametric weighting approach. The effects, which are reported in Appendix A.0.5, then correspond to changes in the probability of being arrested at least once and show a comparable pattern as our main results.

7 | CONCLUSION

Assuming sequential conditional independence, we proposed semi- and nonparametric methods (using either parametric or nonparametric generalized propensity scores) for estimating direct and indirect effects of a continuous treatment based on inverse probability weighting and kernel methods. We demonstrated the asymptotic normality of the estimators under particular regularity conditions and investigated their finite-sample behavior in a simulation study. Finally, we applied the semiparametric method to the Job Corps program. We found this educational intervention to directly and nonlinearly decrease the number of arrests in the fourth year after assignment when controlling for employment as mediator. The semiparametric version of the proposed estimator is available in the "causalweight" package by Bodory and Huber (2018) for the statistical software "R."

As a word of caution, the identifying assumptions considered are rather strong in order to allow for a continuously distributed treatment and possibly multiple mediators with rich support. They may therefore not seem plausible in all settings, in particular when the richness of observed covariates is limited and/or dynamic confounding appears likely. In this case, conditioning on pretreatment covariates is insufficient to control for posttreatment confounders of the mediator–outcome association. In applications where the assumptions seem justifiable, however, the proposed weighting methods are more flexible in terms of modeling assumptions than linear regression-based approaches conventionally used in practice.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [http://qed.econ.queensu.ca/jae/datasets/hsu001/]

REFERENCES

Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business and Economic Statistics*, *33*, 485–505.

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine*, *27*, 1282–1304.

Albert, J. M., & Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, *67*, 1028–1038.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Bia, M., & Mattei, A. (2012). Assessing the effect of the amount of financial aids to piedmont firms using the generalized propensity score. *Statistical Methods and Applications*, *21*, 485–516.

Bodory, H., & Huber, M. (2018). The causalweight package for causal inference in r (SES Working Paper 493). Fribourg, Switzerland: University of Fribourg.

Flores, C. A. (2005). Estimation of dose–response functions and optimal doses with a continuous treatment (Doctoral dissertation). University of California, Berkeley, CA.

Flores, C. A. (2007). Estimation of dose–response functions and optimal doses with a continuous treatment (Working paper). Berkeley, CA: University of California.

Flores, C. A., & Flores-Lagunes, A. (2009). Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness (IZA Discussion Paper No. 4237). Bonn, Germany: Institute of Labor Economics.

Flores, C. A., & Flores-Lagunes, A. (2010). Nonparametric partial identification of causal net and mechanism average treatment effects (Working paper). Gainesville, FL: University of Florida.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., & Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of Job Corps. *Review of Economics and Statistics*, *94*, 153–171.

Frölich, M., & Huber, M. (2017). Direct and indirect treatment effects: causal chains and mediation analysis with instrumental variables. *Journal of Royal Statistical Society, Series B*, *79*, 1645–1666.

Galvao, A. F., & Wang, L. (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, *110*, 1528–1542.

Hayfield, T., & Racine, J. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*, 1–32.

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 73–84). New York, NY: Wiley.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. In *JSM Proceedings, Biometrics Section*, American Statistical Association, Alexandria, VA, pp. 2401–2415.

Horowitz, J. L. (2001). The bootstrap. In Heckman, J. J., & Leamer, E. (Eds.), *Handbook of econometrics* (pp. 3159–3228), Vol. *5*. Amsterdam, Netherlands: Elsevier. https://www.sciencedirect.com/science/article/pii/S157344120105005X

Hsu, Y.-C., Huber, M., & Lai, T.-C. (2018). Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting. *Journal of Econometric Methods*, *8*(1). Advance online publication. https://doi.org/10.1515/jem-2017-0016

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, *29*, 920–943.

Huber, M. (2019). A review of causal mediation analysis for assessing direct and indirect treatment effects (SES Working Paper 500). Fribourg, Switzerland: University of Fribourg.

Huber, M., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, *175*, 1–21.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*, 309–334.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*, 51–71.

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association*, *99*, 854–866.

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, *21*, 141–171.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*, 4–29.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, *5*, 602–619.

Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, *78*, 2021–2042.

Kluve, J., Schneider, H., Uhlendorff, A., & Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society, Series A*, *175*, 587–617.

Lee, Y.-Y. (2018). Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models. arXiv:1811.00157.

Newey, W. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, *10*, 1–21.

Pearl, J. (2001). Direct and indirect effects. In J. Breese, & D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufman.

Petersen, M. L., Sinisi, S. E., & van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, *17*, 276–284.

Powell, J. L., Stock, J. H., & Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, *57*(6), 1403–1430.

Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. In Green, P. J., Hjort, N. L., & Richardson, S. (Eds.), *In highly structured stochastic systems* (pp. 70–81). Oxford, UK: Oxford University Press.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*, 143–155.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, *31*, 161–170.

Schochet, P. Z., Burghardt, J., & Glazerman, S. (2001). National Job Corps study: The impacts of job corps on participants' employment and related outcomes (Report). Washington, DC: Mathematica Policy Research.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does Job Corps work? Impact findings from the national Job Corps study. *American Economic Review*, *98*, 1864–1886.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London, UK: Chapman & Hall.

Smith, J., & Todd, P. (2005). Rejoinder. *Journal of Econometrics*, *125*, 365–375.

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017). Medflex: An R package for flexible mediation analysis using natural effect models. *Journal of Statistical Software*, *76*(11). doi:10.18637/jss.v076.i11

Tchetgen Tchetgen, E. J., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, *40*, 1816–1845.

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., & Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, *63*, 926–934.

Tingley, D., Yamamoto, T., Hirose, K., Imai, K., & Keele, L. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*, 1–38.

van der Weele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, *20*, 18–26.

VanderWeele, T. J., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods*, *2*, 95–115.

Vansteelandt, S., Bekaert, M., & Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, *1*, 129–158.

How to cite this article: Huber M, Hsu Y-C, Lee Y-Y, Lettry L. Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *J Appl Econ*. 2020;35:814–840. https://doi.org/10.1002/jae.2765

APPENDIX A

A.1 | Proof of Theorem 1

Let the supremum norm of a function $A(z)$ be $\|A\| \equiv \sup_z |A(z)|$. Our estimator has the form $\hat{\mathrm{A}}/\hat{B}$. A Taylor expansion gives

$$\frac{\hat{\mathrm{A}}}{\hat{B}} = \frac{A}{B} + \frac{\hat{\mathrm{A}} - A}{B} - \frac{A}{B^2}(\hat{B} - B) + O_p(\|\hat{\mathrm{A}} - A\|\|\hat{B} - B\| + \|\hat{B} - B\|^2). \tag{A1}$$

The numerator of the estimator $\hat{\mu}(d, d)$ is

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\frac{\hat{f}_X(X_i)}{\hat{f}_{DX}(d, X_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\left(\frac{1}{f_D(d|X_i)} + \frac{\hat{f}_X(X_i) - f_X(X_i)}{f_{DX}(d, X_i)} - \frac{\hat{f}_{DX}(d, X_i) - f_{DX}(d, X_i)}{f_D(d|X = X_i) f_{DX}(d, X_i)}\right) \\
&+ O_p\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 K_{2,h_2}^2(D_i - d)\right) O_p\left(\|\hat{f}_{DX} - f_{DX}\|^2\right).
\end{aligned} \tag{A2}$$

The kernel-based estimator satisfies the uniform convergence rate as in lemma B.3 in Newey (1994):

$$\sup_{(d,m,x)\in\mathcal{D}\times\mathcal{M}\times\mathcal{X}} |\hat{f}_{DMX}(d, m, x) - f_{DMX}(d, m, x)| = O_p\left(\left(\frac{\log n}{n{h_1}^s}\right)^{1/2} + {h_1}^{r_1}\right). \tag{A3}$$

Thus the last term in Equation A2 is $O_p\left(h_2^{-1}\left(\left(\log n/(n{h_1}^s)\right)^{-1/2} + {h_1}^{r_1}\right)^2\right) = o_p((nh)^{-1/2})$ by Assumption 3(iv).

We analyze the third term in parentheses of Equation A2:

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i-d)\frac{\hat{f}_{DX}(d,X_i)-f_{DX}(d,X_i)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\\
&=-\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left(\frac{1}{n}\sum_{j=1}^{n}K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)-f_{DX}(d,X_i)\right)\\
&\equiv\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}p(Z_i,Z_j)\\
&=\frac{1}{n}\sum_{i=1}^{n}E[p(Z_i,Z_j)|Z_i]+\frac{1}{n}\sum_{j=1}^{n}E[p(Z_i,Z_j)|Z_j]-E[p(Z_i,Z_j)]+\text{Rem},
\end{aligned}\tag{A4}$$

which is a U-statistic with $Z_i \equiv (Y_i, D_i, X_i)$ and

$$p(Z_i,Z_j)\equiv-\frac{Y_i K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left(K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)-f_{DX}(d,X_i)\right).$$

To control the remainder term Rem, we calculate

$$\begin{aligned}
&E\left[p(Z_i,Z_j)^2\right]\\
&=E\left[\frac{Y_i^2K_{2,h_2}^2(D_i-d)}{f_D^2(d|X=X_i)f_{DX}^2(d,X_i)}E\left[\left(K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)-f_{DX}(d,X_i)\right)^2|Z_i\right]\right]\\
&=O(h_2^{-1}h_1^{-s}).
\end{aligned}$$

Assumption 3(v) implies that $E\left[p(Z_i,Z_j)^2\right]h = O(h_2^{-1}h_1^{-s}h) = o(n)$, which further implies Rem $= o_p((nh)^{-1/2})$ by lemma 3.1 in Powell et al. (1989). The projection $E[p(Z_i,Z_j)|Z_j]$ satisfies

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}E[p(Z_i,Z_j)|Z_j]\\
&=-E\left[\frac{E\left[Y_i|D_i,X_i\right]K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left(\frac{1}{n}\sum_{j=1}^{n}K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)-f_{DX}(d,X_i)\right)|Z_j\right]\\
&=-\frac{1}{n}\sum_{j=1}^{n}\frac{E\left[Y|D=d,X=X_j\right]}{f_D(d|X=X_j)}K_{1,h_1}(D_j-d)+E\left[E\left[Y|D=d,X\right]\right]+O_p({h_2}^{r_2}+{h_1}^{r_1})\\
&=O_p((nh_1)^{-1/2}).
\end{aligned}$$

Also, the projection $E[p(Z_i,Z_j)|Z_i]$ satisfies

$$\begin{aligned}
&E[p(Z_i,Z_j)|Z_i]\\
&=-E\left[\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left(K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)-f_{DX}(d,X_i)\right)|Z_i\right]\\
&=-\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left(E\left[K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)|Z_i\right]-f_{DX}(d,X_i)\right)\\
&=-\frac{Y_iK_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left({h_1}^{r_1}G_i+o_p({h_1}^{r_1})\right),
\end{aligned}$$

where $G_i \equiv \left(\frac{\partial^{r_1}}{\partial d^{r_1}} f_{DX}(d,X_i) + \frac{\partial^{r_1}}{\partial X_i^{r_1}} f_{DX}(d,X_i)\right) \int u^{r_1} K_1(u)du/r_1!$. The last term in Equation A4 is

$$\begin{aligned} E[p(Z_i,Z_j)] &= -E\left[\frac{Y_i K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left(E\left[K_{1,h_1}(D_j-d)K_{1,h_1}(X_j-X_i)|Z_i\right] - f_{DX}(d,X_i)\right)\right] \\ &= -E\left[\frac{Y_i K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left({h_1}^{r_1}G_i + o_p({h_1}^{r_1})\right)\right]. \end{aligned}$$

Therefore

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n E[p(Z_i,Z_j)|Z_i] - E[p(Z_i,Z_j)] &= -\frac{1}{n}\sum_{i=1}^n \frac{Y_i K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left({h_1}^{r_1}G_i + o_p({h_1}^{r_1})\right) \\ &\quad + E\left[\frac{Y_i K_{2,h_2}(D_i-d)}{f_D(d|X=X_i)f_{DX}(d,X_i)}\left({h_1}^{r_1}G_i + o_p({h_1}^{r_1})\right)\right] \\ &= O_p({h_1}^{r_1}/\sqrt{nh_2}) \\ &= o_p((nh)^{-1/2}). \end{aligned}$$

The same argument implies that the second term in the parentheses of Equation A2 is of smaller order. Thus, the asymptotic linear representation for the numerator of $\hat{\mu}(d,d)$ in Equation A2 corresponds to

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n Y_i K_{2,h_2}(D_i-d)\frac{\hat{f}_X(X_i)}{\hat{f}_{DX}(d,X_i)} - E[E[Y|D=d,X]] \\ &= \frac{1}{n}\sum_{i=1}^n \left(Y_i K_{2,h_2}(D_i-d) - E[Y|D=d,X=X_i]K_{1,h_1}(D_i-d)\right)/f_D(d|X_i) + o_p\left((nh)^{-1/2}\right). \end{aligned}$$

The denominator of $\hat{\mu}(d,d)$ is equivalent to the numerator of $\hat{\mu}(d,d)$ by replacing Y_i with 1. By the same argument as above, we obtain

$$\frac{1}{n}\sum_{i=1}^n K_{2,h_2}(D_i-d)\frac{\hat{f}_X(X_i)}{\hat{f}_{DX}(d,X_i)} - 1 = \frac{1}{n}\sum_{i=1}^n \frac{K_{2,h_2}(D_i-d) - K_{1,h_1}(D_i-d)}{f_D(d|X_i)} + o_p\left((nh)^{-1/2}\right).$$

By the Taylor expansion in Equation A1, we then obtain

$$\hat{\mu}(d,d) - \mu(d,d) = \frac{1}{n}\sum_{i=1}^n IF_i + o_p\left((nh)^{-1/2}\right),$$

where $IF_i \equiv (Y_i - \mu(d,d))\frac{K_{2,h_2}(D_i-d)}{f_D(d|X_i)} - (E[Y|D=d,X_i] - \mu(d,d))\frac{K_{1,h_1}(D_i-d)}{f_D(d|X_i)}$. Next we show asymptotic normality by the Lyapounov CLT with third absolute moments. The Lyapounov condition holds because

$$\begin{aligned} &\left(\sum_{i=1}^n \text{var}[IF_i]\right)^{-3/2}\sum_{i=1}^n E\left[|IF_i|^3\right] \\ &= O\left((nh^{-1})^{-3/2}\right)\sum_{i=1}^n E\left[|IF_i|^3\right] = O\left((nh)^{-1/2}\right) = o(1). \end{aligned}$$

By a similar argument, we obtain the asymptotic variance $\lim_{n\to\infty} h\text{var}[IF_i] = V_d$.

Now we turn to $\widehat{\mu}(d,d')$. Let

$$\begin{aligned}\widehat{\Omega}_i &= \widehat{\Omega}(M_i,X_i)\\ &\equiv \frac{\widehat{f}_D(d'|M=M_i,X=X_i)}{\widehat{f}_D(d|M=M_i,X=X_i)\widehat{f}_D(d'|X=X_i)} = \frac{\widehat{f}_{DMX}(d',M_i,X_i)\widehat{f}_X(X_i)}{\widehat{f}_{DMX}(d,M_i,X_i)\widehat{f}_{DX}(d',X_i)}\\ &\equiv \frac{\widehat{A}_i\widehat{F}_i}{\widehat{B}_i\widehat{C}_i} = \Omega_i + \Omega_i\frac{\widehat{A}_i-A_i}{A_i} + \Omega_i\frac{\widehat{F}_i-F_i}{F_i} - \Omega_i\frac{\widehat{B}_i-B_i}{B_i} - \Omega_i\frac{\widehat{C}_i-C_i}{C_i} + O_p\left(\|\widehat{B}_i-B_i\|^2\right).\end{aligned}$$

We use the same argument as in the proof for $\widehat{\mu}(d,d)$ further above. We analyze the numerator of $\widehat{\mu}(d,M(d'))$, $\frac{1}{n}\sum_{i=1}^n Y_iK_{2,h_2}(D_i-d)\widehat{\Omega}_i$. Let *s.o.* stand for smaller order terms. In the *U*-statistic in Equation A4, the *s.o.* are $n^{-1}\sum_{i=1}^n E[p(Z_i,Z_j)|Z_i] - E[p(Z_i,Z_j)] + \text{Rem} = o_p((nh)^{-1/2})$. Thus

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n Y_iK_{2,h_2}(D_i-d)\Omega_i\frac{\widehat{A}_i-A_i}{A_i}\\ &= \frac{1}{n}\sum_{j=1}^n E\left[E[Y_i|D_i,M_i,X_i]K_{2,h_2}(D_i-d)\frac{\Omega_i}{A_i}\left(K_{1,h_1}(D_j-d')K_{1,h_1}(M_j-M_i)K_{1,h_1}(X_j-X_i)-A_i\right)|Z_j\right] + s.o.\\ &= \frac{1}{n}\sum_{j=1}^n E[Y_i|D_i=d,M_i=M_j,X_i=X_j]\frac{\Omega_j}{A_j}f_{DMX}(d,M_j,X_j)K_{1,h_1}(D_j-d')\\ &- E\left[E[Y_i|D_i=d,M_i,X_i]\Omega_i f_{D|MX}(d|M_i,X_i)\right] + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o.\\ &= \frac{1}{n}\sum_{j=1}^n g(d,M_j,X_j)\frac{\Omega_jB_j}{A_j}K_{1,h_1}(D_j-d') - \mu(d,d') + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o.,\end{aligned}$$

where $g(d,M_i,X_i)\equiv E[Y|D=d,M_i,X_i]$. By the same argument, we obtain

$$\begin{aligned}&-\frac{1}{n}\sum_{i=1}^n Y_iK_{2,h_2}(D_i-d)\Omega_i\frac{\widehat{B}_i-B_i}{B_i}\\ &= \frac{1}{n}\sum_{j=1}^n g(d,M_j,X_j)\Omega_jK_{1,h_1}(D_j-d) + \mu(d,d') + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o.,\end{aligned}$$

$$\begin{aligned}&-\frac{1}{n}\sum_{i=1}^n Y_iK_{2,h_2}(D_i-d)\Omega_i\frac{\widehat{C}_i-C_i}{C_i}\\ &= -\frac{1}{n}\sum_{j=1}^n E\left[g(d,M,X_j)|D=d',X=X_j\right]K_{1,h_1}(D_j-d')/f_D(d'|X=X_j)\\ &+ E[Y(d,M(d'))] + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o.,\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n Y_iK_{2,h_2}(D_i-d)\Omega_i\frac{\widehat{F}_i-F_i}{F_i}\\ &= \frac{1}{n}\sum_{j=1}^n E\left[g(d,M,X_j)|D=d',X=X_j\right] - \mu(d,d') + O_p({h_1}^{r_1}+{h_2}^{r_2}) + s.o = O_p(n^{-1/2}).\end{aligned}$$

Collecting all these terms, we obtain the asymptotic linear representation for the numerator $n^{-1}\sum_{i=1}^n Y_iK_{2,h_2}(D_i-d)\widehat{\Omega}_i$. Replacing Y_i with 1 gives the asymptotic linear representation for the denominator: $n^{-1}\sum_{i=1}^n K_{2,h_2}(D_i-d)\widehat{\Omega} = n^{-1}\sum_{i=1}^n\left(K_{2,h_2}(D_i-d)-K_{1,h_1}(D_i-d)\right)\Omega_i + o_p((nh)^{-1/2})$. The Lyapounov CLT gives the asymptotic normality.

A.2 | Proof of Theorem 2

We consider the estimator $\widehat{\mu}(d,d)$. Let $\Omega_i(\gamma) = 1/f_D(d|X_i)$ and $\widehat{\Omega}_i(\gamma) = 1/f_D(d|X_i;\widehat{\gamma}_x)$. By a mean-value expansion, it holds that $\widehat{\Omega}_i(\gamma) - \Omega_i(\gamma) = -\bar{w}_i^{-2}(f_D(d|X_i) - f_D(d|X_i;\widehat{\gamma}_x))$ for some $\bar{w}_i$ between $f_D(d|X_i)$ and $f_D(d|X_i;\widehat{\gamma}_x)$. Then $\widehat{\Omega}_i(\gamma) - \Omega_i(\gamma) = O_p(n^{-1/2})$ uniformly over i. We start with the numerator of the estimator $\widehat{\mu}(d,d)$. Note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\widehat{\Omega}_i(\gamma) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\Omega_i(\gamma) + \frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)(\widehat{\Omega}_i(\gamma) - \Omega_i(\gamma)) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\Omega_i(\gamma) + O_p((nh_2)^{-1/2})O_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i K_{2,h_2}(D_i - d)\Omega_i(\gamma) + o_p(1),
\end{aligned}$$

where the second equality holds by a similar argument as in theorem 2 of Abrevaya, Hsu, and Lieli (2015). The derivation for the denominator follows the same arguments. By the Taylor expansion (Equation A1) and $E[\Omega|D = d]f_D(d) = 1$,

$$\widehat{\mu}(d,d) - \mu(d,d) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \mu(d,d)}{f_D(d|X_i)}\right)K_{2,h_2}(D_i - d) + O_p\left((nh_2)^{-1}\right).$$

The asymptotic normality is shown by the Lyapounov CLT with third absolute moments as the arguments in the proof of Theorem 1. The proof for $\widehat{\mu}(d,d')$ is analogous and therefore omitted.

A.3 | Results for nontreatement $d' = 0$

A.4 | Results using generalized additive regression models for M and Y with $d' = 40$

A.5 | Results for binary outcome with $d' = 40$

A.6 | Description of the categorical variables

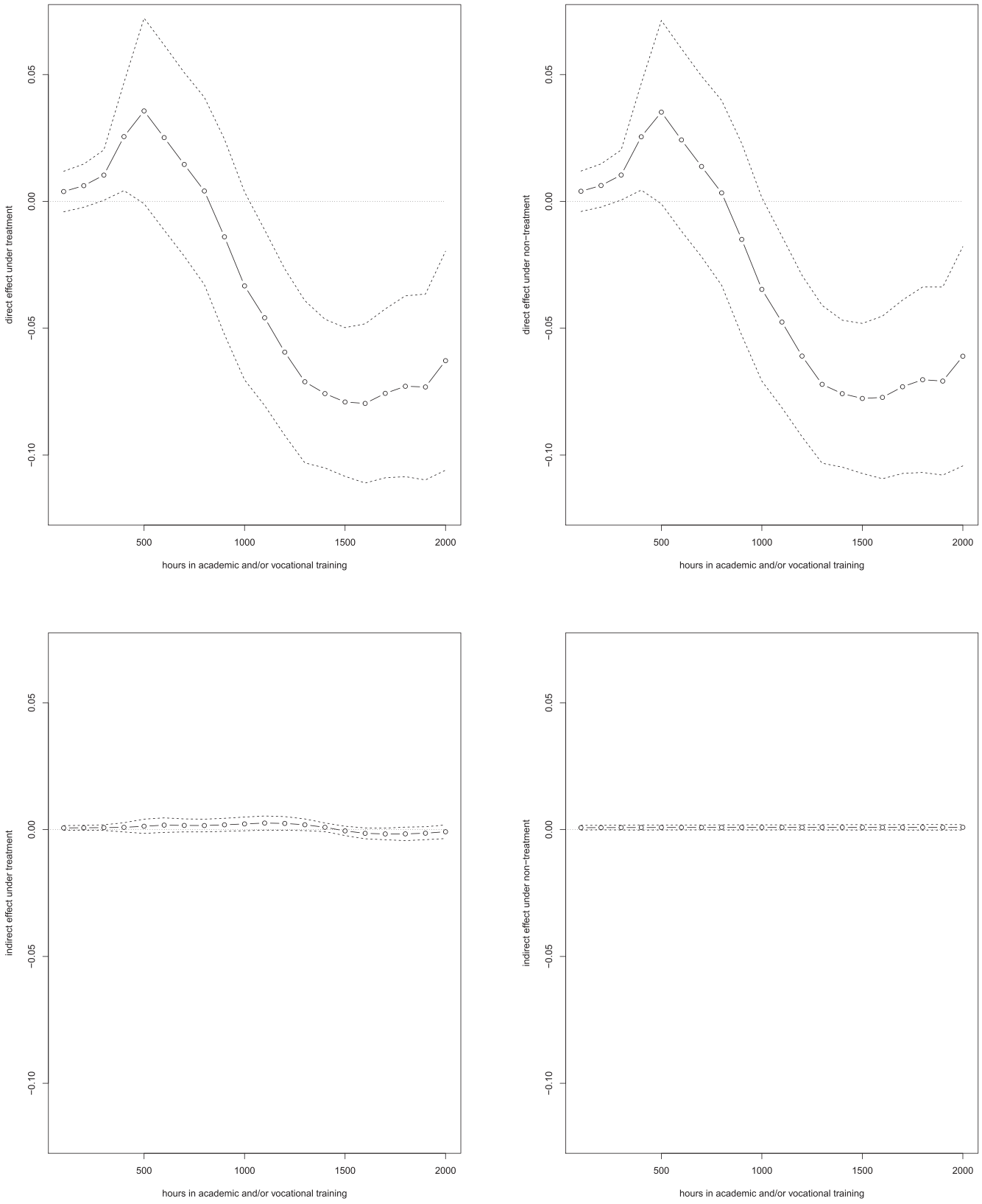

FIGURE A1 Direct effects $\widehat{\theta}_{d,0}(d)$ (top left) and $\widehat{\theta}_{d,0}(0)$ (top right) as well as indirect effects $\widehat{\delta}_{d,0}(d)$ (bottom left) and $\widehat{\delta}_{d,0}(0)$ (bottom right) for $d \in \{100, 200, \ldots, 1,900, 2,000\}$

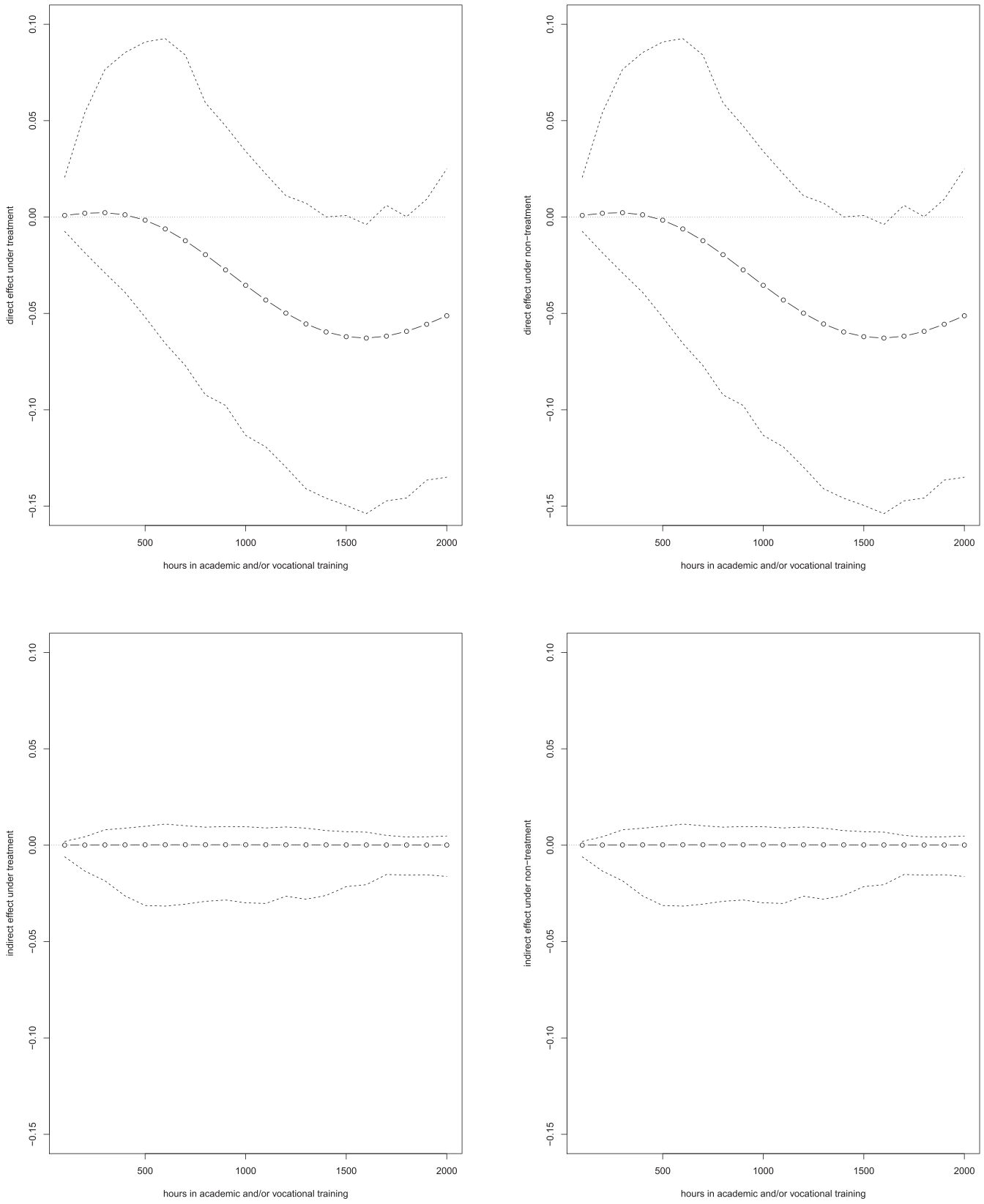

FIGURE A2 Direct effects $\widehat{\theta}_{d,40}(d)$ (top left) and $\widehat{\theta}_{d,40}(40)$ (top right) as well as indirect effects $\widehat{\delta}_{d,40}(d)$ (bottom left) and $\widehat{\delta}_{d,40}(40)$ (bottom right) for $d \in \{100, 200, \ldots, 1,900, 2,000\}$

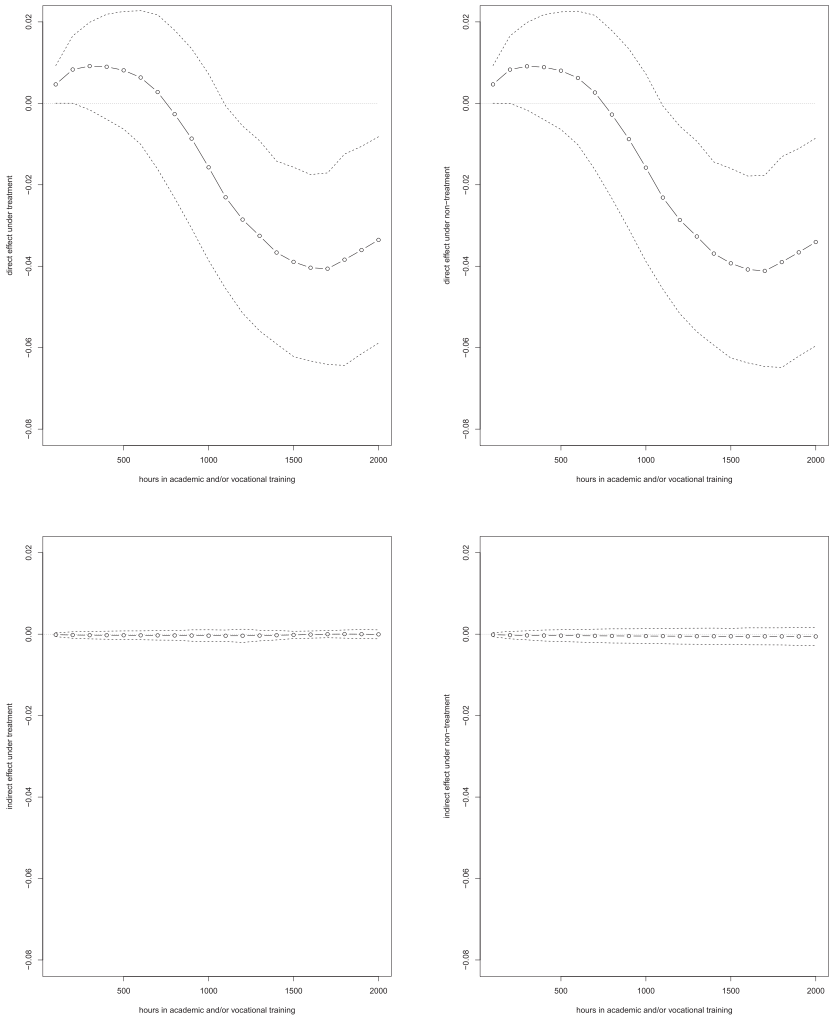

FIGURE A3 Direct effects $\widehat{\theta}_{d,40}(d)$ (top left) and $\widehat{\theta}_{d,40}(40)$ (top right) as well as indirect effects $\widehat{\delta}_{d,40}(d)$ (bottom left) and $\widehat{\delta}_{d,40}(40)$ (bottom right) for $d \in \{100, 200, \ldots, 1,900, 2,000\}$

TABLE A1 Description table of the categorical variables

Variable	Description	Categories
total household gross income	Total income (in USD) from	1 lower than 3,000, 2 if between 3,000 and 6“000, 3 if between 6,000 and 9,000, 4
	all household members last year	if between 9,000 and 12,000, 5 if between 12,000 and 15,000, 6 if between 15,000
		and 18,000, 7 if larger than 18,000;
total personal gross income	Total personal income (in USD) last year	1 lower than 3,000, 2 if between 3,000 and 6,000, 3 if between 6,000 and 9,000, 4
		if between 9,000 and 12,000, 5 if between 12,000 and 15,000, 6 if between 15,000
		and 18,000, 7 if larger than 18,000;
welfare receipt during childhood	How often got welfare while growing up	1 if never, 2 if occasionally, 3 if about half the time, 4 if most or all of the time;
extent of marijuana use	How often used marijuana in last year	1 if daily, 2 if a few times each week, 3 if a few times each month, 4 if less often than that;
extent of hallucinogen use	How often used LSD/peyote/psilocybin/other	1 if daily, 2 if a few times each week, 3 if a few times each month, 4 if less often than that;
	hallucinogenic drugs last year	
extent of smoking	How often smoked last year	1 if daily, 2 if a few times each week, 3 if a few times each month, 4 if less often than that;
extent of alcohol consumption	How often used alcohol last year	1 if daily, 2 if a few times each week, 3 if a few times each month, 4 if less often than that;
time spent by recruiter speaking of Job Corps	How much time did the recruiter	1 if one hour or less, 2 if between 1 and 2 hours, 3 if between 2 and 3 hours,
	spend with you talking about JC	4 if more than 3 hours;
extent of recruiter support	How much did the recruiter encourage you	1 if encouraged a lot, 2 if encouraged a little, 3 if discouraged a little, 4 if discouraged
		a lot, 5 if offered no encouragement or did not express an opinion;
expected improvement in maths	Extent to which JC will help with math skills (expectation)	1 if a lot, 2 if a little, 3 if not at all;
expected improvement in reading skills	Extent to which JC will help with reading skills (expectation)	1 if a lot, 2 if a little, 3 if not at all;
expected improvement in social skills	Extent to which JC will help with social skills (expectation)	1 if a lot, 2 if a little, 3 if not at all;
expected to be training for a job	Extent to which JC will help to provide	1 if a lot, 2 if a little, 3 if not at all.
	training for a specific job (expectation)	