HETEROGENEOUS COEFFICIENTS, CONTROL VARIABLES, AND IDENTIFICATION OF MULTIPLE TREATMENT EFFECTS

WHITNEY K. NEWEY† AND SAMI STOULI§

Abstract. Multidimensional heterogeneity and endogeneity are important features of models with multiple treatments. We consider a heterogeneous coefficients model where the outcome is a linear combination of dummy treatment variables, with each variable representing a different kind of treatment. We use control variables to give necessary and sufficient conditions for identification of average treatment effects. With mutually exclusive treatments we find that, provided the heterogeneous coefficients are mean independent from treatments given the controls, a simple identification condition is that the generalized propensity scores (Imbens, 2000) be bounded away from zero and that their sum be bounded away from one, with probability one. Our analysis extends to distributional and quantile treatment effects, as well as corresponding treatment effects on the treated. These results generalize the classical identification result of Rosenbaum and Rubin (1983) for binary treatments.

Keywords: Treatment effect; Multiple treatments; Heterogeneous coefficients; Control variable; Identification; Conditional nonsingularity; Propensity score.

1. Introduction

Models that allow for multiple treatments are important for program evaluation and the estimation of treatment effects (Cattaneo, 2010; Imai and van Dyk, 2004; Imbens, 2000; Graham and Pinto, 2018; Lechner, 2001). A general class is heterogeneous coefficients models where the outcome is a linear combination of dummy treatment variables and unobserved heterogeneity. These models allow for multiple treatment regimes, with each dummy variable representing a different kind of treatment. These models also feature multidimensional heterogeneity, with the dimension of unobserved heterogeneity being determined by the number of treatment regimes.

† Department of Economics, MIT, wnewey@mit.edu.
§ Department of Economics, University of Bristol, s.stouli@bristol.ac.uk.

Endogeneity is often a problem in these models because we are interested in the effect of treatment variables on an outcome, and the treatment variables are correlated with heterogeneity. Control variables provide an important means of controlling for endogeneity with multidimensional heterogeneity. For treatment effects, a control variable is an observed variable that makes heterogeneity and treatment variables independent when it is conditioned on (Rosenbaum and Rubin, 1983).

We use control variables to give necessary and sufficient conditions for identification of average treatment effects based on conditional nonsingularity of the second moment matrix of the vector of dummy treatment variables given the controls. This first main result is familiar in the binary treatment case, but its generalization to multiple treatments appears to be new. With mutually exclusive treatments we find that, provided the heterogeneous coefficients are mean independent from treatments given the controls, a simple identification condition is that the generalized propensity scores (Imbens, 2000) be bounded away from zero and that their sum be bounded away from one, with probability one. This condition is the same as common support, that the support of treatment variables conditional on the controls is equal to the marginal support of the treatment variables. Thus our second main contribution is to show that, with mutually exclusive treatments, conditional mean independence and common support are jointly sufficient for identification, a substantial weakening of the standard assumption that conditional independence hold jointly with common support (e.g., Frölich, 2004). We also extend our analysis to distributional and quantile treatment effects, as well as corresponding treatment effects on the treated. These results provide an important generalization of Rosenbaum and Rubin (1983)'s classical identification result for binary treatments.

2. Modeling of Treatment Effects

2.1. **Modeling framework.** Let Y denote an outcome variable of interest, and D a vector of dummy variables $D(t)$, $t \in \mathcal{T} \equiv \{1, \ldots, T\}$, taking value one if treatment t occurs and zero otherwise, and β a structural disturbance vector of finite dimension. We consider a heterogeneous coefficients model of the form

$$Y = p(D)^{\mathrm{T}}\beta, \qquad p(D) = \{1, D(1), \ldots, D(T)\}^{\mathrm{T}}. \tag{1}$$

This model is linear in the treatment dummy variables, with coefficients β that are random and need not be independent of D.

When the potential outcome framework of Rubin (1974) is extended to mutually exclusive treatment regimes in the definition of D, linearity in model (1) arises naturally. Denote the vector of potential outcomes by $\{Y(1), \ldots, Y(T)\}^{\mathrm{T}}$, and the potential outcome in the absence of treatment by $Y(0)$. The observed outcome Y and the vector of potential outcomes are related by

$$Y = Y(0) + \sum_{t=1}^{\mathrm{T}} D(t)\{Y(t) - Y(0)\},$$

which is of the form (1) upon setting $\beta = (\beta_0, \beta_1, \ldots, \beta_T)^{\mathrm{T}}$ with $\beta_0 \equiv Y(0)$ and $\beta_t \equiv Y(t) - Y(0)$, $t \in \mathcal{T}$.

Mutually exclusive treatment regimes are important in a wide variety of nonexperimental settings, such as program or policy evaluation with a multivalued treatment (e.g., Ao et al., 2021; Lechner, 2002; Uysal, 2015). Consider for example evaluation of active labor market programs with a treatment taking on $T \geq 2$ values, according to different types or levels of program participation. Central objects of interest are average effects on earnings for each treatment value,

$$E(\beta_t) = E\{Y(t) - Y(0)\}, \quad t \in \mathcal{T}.$$

Allowing for multivalued treatments thus permits to capture different average effects across program types or levels, going beyond the sole effect of program participation considered in binary treatment analysis. Another important example is policy or medical treatment evaluation with non-mutually exclusive policies or treatments, implemented both separately and jointly. In that case, a distinct treatment dummy variable is assigned to each policy or treatment and to each implemented policy mix (Becker and Egger, 2013; Tortú et al., 2020) or combined therapy (Feng et al., 2012; Nian et al., 2019). Resulting treatment regimes in the definition of D are then mutually exclusive, by construction.

A main motivation for defining the components of D according to mutually exclusive treatment regimes is the general validity of model (1) in that case. In contrast, when treatment regimes are non-mutually exclusive, model (1) restricts the average effect of any combination $\mathcal{C}$ of $K \leq T$ treatments $t_1, \ldots, t_K \in \mathcal{T}$ to be additive in the average effects of each component of $\mathcal{C}$, i.e., the average treatment effect of $\mathcal{C}$ is

$$\sum_{t \in \mathcal{C}} E(\beta_t) = \sum_{t \in \mathcal{C}} E\{Y(t) - Y(0)\}, \quad \mathcal{C} = \{t_1, \ldots, t_K\}. \tag{2}$$

In addition to average effects of each treatment, model (1) is then able to capture average effects of potentially complex interventions by restricting the form the average effects of combined treatments can take. The additivity restriction has been used in evaluation of randomized medical experiments implementing combinations of a large number of treatments (see Petropoulou et al., 2021, for a literature review), but does not appear to be common in nonexperimental settings.

In general, heterogeneous coefficients β need not be independent of D because of confounding factors denoted X. Here we assume that these factors are observable and that there is sufficient independent variation in D from β once conditioning on X. In the empirical examples above, X includes a variety of individual characteristics of program participants, such as age, gender, measures of cognitive and non-cognitive skills, as well as socio-economic characteristics. Formally, we assume that the vector β is mean independent of the endogenous treatments D, conditional on an observable control variable X.

Assumption 1. For the model in (1), there exists a control variable X such that $E(\beta \mid D, X) = E(\beta \mid X)$.

The Rosenbaum and Rubin (1983) treatment effects model is included as a special case where $D \in \{0, 1\}$ is a treatment dummy variable that is equal to one if treatment occurs and equals zero without treatment, and

$$p(D) = (1, D)^{\mathrm{T}}.$$

In this case $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$ is two dimensional with β_0 giving the outcome without treatment, and β_1 being the treatment effect. Here the control variables in X would be observable variables such that Assumption 1 holds, i.e., the coefficients (β_0, β_1) are mean independent of treatment conditional on controls; this is the unconfoundedness assumption of Rosenbaum and Rubin (1983).

2.2. **The average structural function.** A central object of interest in model (1) is the average structural function given by $\mu(D) \equiv p(D)^{\mathrm{T}} E(\beta)$; see Chamberlain (1984), Blundell and Powell (2003) and Wooldridge (2005). This function is also referred to as the dose-response function in the statistics literature (e.g., Imbens, 2000). When $D \in \{0, 1\}$ is a dummy variable for treatment, $\mu(0)$ gives the average outcome if every unit remained untreated and $\mu(1)$ the average outcome if every

unit were treated, with $\mu(1) - \mu(0)$ being the average treatment effect. In general, the average effect of some treatment $t \in \mathcal{T}$ is

$$\mu(e_t) - \mu(0_T),$$

with $e_t = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$ defined as a T-vector with all components equal to zero, except the tth, which is one, and 0_T a T-vector of zeros. Pairwise average treatment effect comparisons are formed as $\mu(e_t) - \mu(e_s)$, for any $s, t \in \mathcal{T}$, $s \neq t$. For non-mutually exclusive treatment regimes, the average effect of some combination $\mathcal{C}$ of treatments $t_1, \ldots, t_K \in \mathcal{T}$, $K \leq T$, is formed as $\sum_{s \in \mathcal{C}} \{\mu(e_s) - \mu(0_T)\}$, $\mathcal{C} = \{t_1, \ldots, t_K\}$, and the corresponding relative average effect with respect to some treatment $t \in \mathcal{T}$ as $\sum_{s \in \mathcal{C}} \{\mu(e_s) - \mu(e_t)\}$.

The conditional mean independence assumption and the form of the structural function $p(D)^{\mathrm{T}}\beta$ in (1) together imply that the control regression function of Y given (D, X), $E(Y \mid D, X)$, is a linear combination of the treatment variables:

(3)
$$E\left(Y|D, X\right) = p(D)^{\mathrm{T}} E\left(\beta|D, X\right) = p(D)^{\mathrm{T}} E\left(\beta|X\right) = p(D)^{\mathrm{T}} q_0(X), \; q_0(X) \equiv E\left(\beta|X\right).$$

The average structural function can thus be expressed as a known linear combination of $E\{q_0(X)\}$ from equation (3). By iterated expectations,

$$p(D)^{\mathrm{T}} E\left\{q_0(X)\right\} = p(D)^{\mathrm{T}} E\left\{E\left(\beta \mid X\right)\right\} = \mu(D). \tag{4}$$

We use the varying coefficient structure of the control regression function (3) and the implied linear form of $\mu(D)$ to give conditions that are necessary as well as sufficient for identification. For non-mutually exclusive treatment regimes, Appendix A gives an example of a model for which the implied average structural function is of the linear form (4) while the average effect of some combination of treatments takes the additive form (2).

3. Identification Analysis

3.1. **Main results.** Under the maintained Assumption 1, a sufficient condition for identification of the average structural function is nonsingularity of the second moment matrix of the treatment dummies given the controls,

$$E\left\{p(D)p(D)^{\mathrm{T}} \mid X\right\},$$

with probability one. Under the additional assumption that $E\{p(D)p(D)^{\mathrm{T}}\}$ is nonsingular, this condition is also necessary.

Theorem 1 states our first main result. The proofs of all formal results are given in Appendix B.

Theorem 1. *Suppose that $E(\|\beta\|^2) < \infty$, $E\{p(D)p(D)^T\}$ is nonsingular, and Assumption 1 holds. Then: $E\{p(D)p(D)^T \mid X\}$ is nonsingular with probability one if, and only if, $\mu(D)$ is identified.*

When $D \in \{0, 1\}$ and $p(D) = (1, D)^{\mathrm{T}}$, the identification condition becomes the standard condition for the treatment effect model

$$Y = \beta_0 + \beta_1 D, \quad E(\beta \mid D, X) = E(\beta \mid X), \quad \beta \equiv (\beta_0, \beta_1)^{\mathrm{T}}.$$

The identification condition is that the conditional second moment matrix of $(1, D)^{\mathrm{T}}$ given X is nonsingular with probability one, which is the same as

$$\mathrm{var}(D \mid X) = P(X)\{1 - P(X)\} > 0, \quad P(X) \equiv \Pr(D = 1 \mid X), \tag{5}$$

with probability one, where $P(X)$ is the propensity score. Here we can see that the identification condition is the same as $0 < P(X) < 1$ with probability one, which is the standard identification condition.

Because $p(D)$ includes an intercept, the identification condition is the same as nonsingularity of the variance matrix $\mathrm{var}(D \mid X)$ with probability one. This result generalizes (5).

Theorem 2. *$E\{p(D)p(D)^T \mid X\}$ is nonsingular with probability one if, and only if, the variance matrix $var(D \mid X)$ is nonsingular with probability one.*

Considerable simplification occurs with mutually exclusive treatment regimes, which allows for the formulation of an equivalent condition for nonsingularity of $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ solely in terms of the generalized propensity scores (Imbens, 2000). This result generalizes the standard identification condition for binary D.

Theorem 3. *With mutually exclusive treatment regimes, $E\{p(D)p(D)^T \mid X\}$ is nonsingular with probability one if, and only if, $\Pr\{D(t) = 1 \mid X\} > 0$ for each $t \in \mathcal{T}$ and*

$$\Sigma_{s=1}^{T} \Pr\{D(s) = 1 \mid X\} < 1,$$

with probability one.

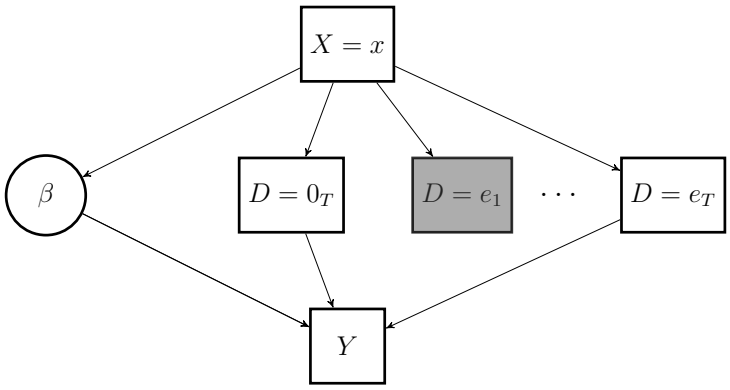

FIGURE 1. A directed acyclic graph representation of lack of identification caused by conditional singularity: treatments are mutually exclusive and treatment 1 occurs with zero probability given $X = x$, and hence the corresponding treatment effect is not identified.

For mutually exclusive treatment regimes, the two standard assumptions for identification are common support, i.e., $\Pr(D = 0_T \mid X) > 0$ and $\Pr\{D(t) = 1 \mid X\} > 0$ with probability one for each $t \in \mathcal{T}$, and conditional independence, i.e.,

$$Y(t) \perp D \mid X, \quad (t = 0, 1, \ldots, T); \tag{6}$$

cf., for instance, Frölich (2004, pp. 190–192) for a review. By conditional probabilities adding up to unity, Theorem 3 shows that common support is equivalent to conditional nonsingularity, and hence is necessary as well as sufficient for identification under Assumption 1. It follows that, provided common support holds, identification only requires conditional mean independence, and assumption (6) is not necessary.

For non-mutually exclusive treatment regimes, conditional independence assumption (6) and common support are not jointly necessary either. In that case, the marginal support $\mathcal{D}$ of D has cardinality $\widetilde{T} > T$. Suppose there exist both a subset $\widetilde{\mathcal{D}} \subset \mathcal{D}$ of cardinality T such that $E\{\mathbb{1}(D \in \widetilde{\mathcal{D}})p(D)p(D)^{\mathrm{T}} \mid X\}$ is nonsingular with probability one, and a value $\overline{d} \in \mathcal{D}\backslash\widetilde{\mathcal{D}}$ such that $\Pr(D = \overline{d} \mid X) = 0$ with positive probability. Then common support does not hold but $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is nonsingular with probability one, and hence $\mu(D)$ is identified under Assumption 1. Therefore, Theorems 1 and 3 together establish that conditional independence assumption (6) and common support are not jointly necessary for identification, for either type of treatment regime.

Theorem 3 also clarifies the role played by conditional nonsingularity in identification. For mutually exclusive treatment regimes $\Pr\{D(t) = 1 \mid X\} = \Pr(D = e_t \mid X)$, $t \in \mathcal{T}$, and hence conditional singularity on a set with positive probability means that at least one of the events $\{D = 0_T\}$,$\{D = e_1\}$,...,$\{D = e_T\}$, has probability zero conditional on X on that set. Therefore, in this instance, failure of identification occurs because common support does not hold. Using a directed acyclic graph (Pearl, 2009), Figure 1 illustrates this failure with the event $\{D = e_1\}$ having probability zero given $X = x$, for almost every x in a set with positive probability. The absence of an arrow between this event and Y reflects failure of identification.

3.2. **Extensions.** Our identification results are useful for the analysis of other interesting objects. When Assumption 1 is strengthened to conditional independence

$$\beta \perp D \mid X, \tag{7}$$

conditional nonsingularity is also sufficient for identification of distributional and quantile treatment effects. Define the distribution structural function $G(y, d)$ and, when Y is continuous, the quantile structural function $Q(\tau, d)$ by

$$G(y, d) \equiv \Pr\{p(d)^{\mathrm{T}}\beta \leq y\}, \quad Q(\tau, d) \equiv \tau^{\text{th}} \text{ quantile of } p(d)^{\mathrm{T}}\beta,$$

where d is fixed in these expressions. Distributional and quantile treatment effects are formed as $G(y, e_t) - G(y, 0_T)$ and $Q(\tau, e_t) - Q(\tau, 0_T)$, respectively, for each $t \in \mathcal{T}$, and pairwise distributional and quantile treatment comparisons as $G(y, e_t) - G(y, e_s)$ and $Q(\tau, e_t) - Q(\tau, e_s)$, respectively, for any $s, t \in \mathcal{T}$, $s \neq t$.

With mutually exclusive treatment regimes, by Theorem 3 the conditional support of D given X coincides with the marginal support of D, and hence the conditional support of X given D coincides with the marginal support of X, with probability one. Therefore, by Imbens and Newey (2009, p. 1489) conditional nonsingularity and conditional independence property (7) together imply identification of $G(Y, D)$ and, when Y is continuous, also of $Q(\tau, D)$, from

$$G(Y, D) = \int F_{Y|DX}(Y \mid D, X = x) F_X(dx), \quad Q(\tau, D) = G^{-1}(\tau, D),$$

where $F_{Y|DX}(Y \mid D, X)$ and $F_X(X)$ are the cumulative distribution functions of Y given (D, X) and of X, respectively, and $\tau \mapsto G^{-1}(\tau, D)$ denotes the inverse function of $y \mapsto G(y, D)$.

With non-mutually exclusive treatment regimes, conditional nonsingularity need not coincide with common support. Identification without common support can nonetheless be achieved under additional restrictions imposed on model (1). When Y is continuous and letting $Q_{\beta_t|DX}(u \mid D, X)$ denote the conditional quantile function of β_t given (D, X), $u \in (0, 1)$, an example of sufficient model restrictions is that unobserved heterogeneity components β_t satisfy conditional independence property (7) as well as the additional scalar heterogeneity restriction

$$\beta_t = Q_{\beta_t|DX}(U \mid D, X), \quad U \mid D, X \sim \text{Un}(0, 1), \quad (t = 0, 1, \ldots, T), \tag{8}$$

where the unobservable U is the same for each β_t. The control quantile regression function of Y given (D, X) then takes the linear form

$$\begin{aligned} Q_{Y|DX}(U \mid D, X) &= p(D)^{\mathrm{T}} q_U(X), \\ q_U(X) &\equiv \{Q_{\beta_0|X}(U \mid X), Q_{\beta_1|X}(U \mid X), \ldots, Q_{\beta_T|X}(U \mid X)\}^{\mathrm{T}}, \end{aligned}$$

by strict monotonicity of $u \mapsto Q_{\beta_0|DX}(u \mid D, X) + \sum_{t=1}^{T} D(t) Q_{\beta_t|DX}(u \mid D, X)$. Conditional nonsingularity implies identification of $q_U(X)$, and hence of $Q_{Y|DX}(U \mid D, X)$. Since the structural functions $G(Y, D)$ and $Q(\tau, D)$ are known functionals of $F_{Y|DX}(Y \mid D, X)$, the relation

$$F_{Y|DX}(Y \mid D, X) = \int_0^1 \mathbb{1}\{Q_{Y|DX}(u \mid D, X) \leq Y\} du$$

implies identification of distributional and quantile treatment effects (Newey and Stouli, 2021).

Theorem 4 summarizes the above discussion of the role of conditional nonsingularity in identification of distributional and quantile treatment effects.

Theorem 4. *Suppose that conditional independence property (7) holds and $E\{p(D)p(D)^T \mid X\}$ is nonsingular with probability one. The following hold: (i) with mutually exclusive treatment regimes, $G(Y, D)$ and, when Y is continuous, also $Q(\tau, D)$, $\tau \in (0, 1)$, are identified; (ii) with non-mutually exclusive treatment regimes and continuous outcome Y, if the scalar heterogeneity restriction (8) holds and $\sup_{u \in (0,1)} E\{||q_u(X)||^2\} < \infty$, then $G(Y, D)$ and $Q(\tau, D)$, $\tau \in (0, 1)$, are identified.*

Other objects of interest include treatment effects on the treated. For some specified treatment $s \in \mathcal{T}$, average effects are formed using the average structural function

for the treated, $\mu(D, e_s) \equiv p(D)^{\mathrm{T}} E(\beta \mid D = e_s)$. Distributional and, when Y is continuous, quantile treatment effects are formed using the distribution and quantile structural functions for the treated,

$$\begin{aligned} G(y, d, e_s) &\equiv \Pr\{p(d)^{\mathrm{T}}\beta \leq y \mid D = e_s\}, \\ Q(\tau, d, e_s) &\equiv \tau^{\text{th}} \text{ quantile of } p(d)^{\mathrm{T}}\beta \text{ given } D = e_s, \end{aligned}$$

respectively, where d is fixed in these expressions. These structural objects are useful for decomposition and counterfactual analysis (e.g., Ao et al., 2021). The average effect of treatment t on units treated with treatment s is $\mu(e_t, e_s) - \mu(0_T, e_s)$, and distributional and quantile effects of treatment t on units treated with treatment s are $G(y, e_t, e_s) - G(y, 0_T, e_s)$ and $Q(\tau, e_t, e_s) - Q(\tau, 0_T, e_s)$, respectively.

Let $\mathcal{X}(s)$ denote the conditional support of X given $D = e_s$. The average structural function for the treated can be expressed as a linear combination of $E\{q_0(X) \mid D = e_s\}$. By conditional mean independence and iterated expectations,

$$p(D)^{\mathrm{T}} E\{q_0(X) \mid D = e_s\} = p(D)^{\mathrm{T}} E\{E(\beta \mid X, D = e_s) \mid D = e_s\} = \mu(D, e_s),$$

and hence $\mu(D, e_s)$ is identified if $q_0(X)$ is identified on $\mathcal{X}(s)$. Thus, for average treatment effects on the treated, the identification condition becomes nonsingularity of $E\{p(D)p(D)^{\mathrm{T}} \mid X = x\}$ for almost every x in the set $\mathcal{X}(s)$. This conditional nonsingularity condition is also necessary for identification of $\mu(D, e_s)$ under the additional condition that $E\{p(D)p(D)^{\mathrm{T}}\}$ is nonsingular.

Theorem 5. *Suppose that Assumption 1 holds, $E\{p(D)p(D)^{T}\}$ is nonsingular, and $\sup_{x \in \mathcal{X}(s)} E(\|\beta\|^2 \mid X = x) < \infty$ for some specified $s \in \mathcal{T}$ such that $\Pr\{\mathcal{X}(s)\} > 0$. Then: $E\{p(D)p(D)^{T} \mid X = x\}$ is nonsingular for almost every $x \in \mathcal{X}(s)$ if, and only if, $\mu(D, e_s)$ is identified. Furthermore, with mutually exclusive treatment regimes and for almost every $x \in \mathcal{X}(s)$, $E\{p(D)p(D)^{T} \mid X = x\}$ is nonsingular if, and only if, $\Pr\{D(t) = 1 \mid X = x\} > 0$ for each $t \in \mathcal{T}$ and $\Sigma_{s=1}^{T} \Pr\{D(s) = 1 \mid X = x\} < 1$.*

If conditional independence property (7) holds, the distribution and, when Y is continuous, quantile structural functions for the treated also are identified, from

$$G(Y, D, e_s) = \int F_{Y|DX}(Y|D, X = x) F_{X|D}(dx|D = e_s), \; Q(\tau, D, e_s) = G^{-1}(\tau, D, e_s),$$

respectively, where $\tau \mapsto G^{-1}(\tau, D, e_s)$ denotes the inverse function of $y \mapsto G(y, D, e_s)$. Here identification only requires the support of X conditional on D to contain $\mathcal{X}(s)$

with probability one, and hence that the support of D conditional on $X = x$ be the same as the marginal support of D for almost every $x \in \mathcal{X}(s)$. With mutually exclusive treatment regimes, this support condition is equivalent to nonsingularity of $E\{p(D)p(D)^{\mathrm{T}} \mid X = x\}$ for almost every $x \in \mathcal{X}(s)$, by Theorem 5. Therefore, this conditional nonsingularity condition is sufficient for identification. With non-mutually exclusive treatment regimes, this condition is also sufficient for identification of $q_U(X)$ on $\mathcal{X}(s)$, and hence of $Q_{Y|DX}(U \mid D, X)$ and $F_{Y|DX}(Y \mid D, X)$ on $\mathcal{D} \times \mathcal{X}(s)$, when the outcome Y is continuous and the scalar heterogeneity restriction (8) holds. Thus, results analogous to Theorem 4 hold for distribution and quantile treatment effects on the treated.

Theorem 6. *Suppose that conditional independence property (7) holds and $E\{p(D)p(D)^{T}|X = x\}$ is nonsingular for almost every $x \in \mathcal{X}(s)$, for some specified $s \in \mathcal{T}$ such that $\Pr\{\mathcal{X}(s)\} > 0$. The following hold: (i) with mutually exclusive treatment regimes, $G(Y, D, e_s)$ and, when Y is continuous, also $Q(\tau, D, e_s)$, $\tau \in (0, 1)$, are identified; (ii) with non-mutually exclusive treatment regimes and continuous outcome Y, if the scalar heterogeneity restriction (8) holds and $\sup_{(u,x)\in(0,1)\times\mathcal{X}(s)} E\{||q_u(X)||^2 \mid X = x\} < \infty$, then $G(Y, D, e_s)$ and $Q(\tau, D, e_s)$, $\tau \in (0, 1)$, are identified.*

4. Discussion

The heterogeneous coefficients formulation we propose for multiple treatment effects reveals the central role of the conditional nonsingularity condition for identification. Because this condition is in principle testable, establishing that it is also necessary demonstrates testability of identification (e.g., Breusch, 1986). With mutually exclusive treatments, the formulation of the equivalent common support condition in Theorem 3 thus relates testability of identification to the generalized propensity scores. This is a generalization of the relationship between testability of identification and the propensity score in the binary treatment case.

Conditions that are both necessary and sufficient are also important for the determination of minimal conditions for identification. In an unpublished 2004 working paper (cemmap CWP03/04), Wooldridge considers a restricted version of our model with $E(D \mid \beta, X) = E(D \mid X)$ and $E\{p(D)p(D)^{\mathrm{T}} \mid \beta, X\} = E\{p(D)p(D)^{\mathrm{T}} \mid X\}$,

and shows that $q_0(X)$ is identified if $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is invertible. The additional conditional second moments assumption implies that his identification condition differs from ours. Thus his result and proof do not apply in our setting which only assumes conditional mean independence $E(\beta \mid D, X) = E(\beta \mid X)$, and our results show that conditional second moments independence is not necessary for identification in multiple treatment effect models. Graham and Pinto (2018) consider a related approach in work independent of the first version of this paper (Newey and Stouli, 2018) where we derived our identification result (Lemma 1 in the Appendix). The conditional nonsingularity condition we propose is weaker than their identification condition, and we study necessity as well as sufficiency for identification of average treatment effects.

We analyze the role of conditional nonsingularity for identification of multiple treatment effects under the maintained conditional mean independence Assumption 1. Although itself not testable in general, this assumption is substantially weaker than the standard conditional independence property (6). In the general case of mutually exclusive treatment regimes, the relaxation of conditional independence afforded by our heterogeneous coefficients approach is the same as the conditional mean independence condition

$$E\{Y(t) \mid D, X\} = E\{Y(t) \mid X\} \quad (t = 0, 1, \ldots, T), \tag{9}$$

because the formulation of Assumption 1 in terms of potential outcomes,

$$E\{Y(0)|D, X\} = E\{Y(0)|X\}, \;\; E\{Y(t)-Y(0)|D, X\} = E\{Y(t)-Y(0)|X\}, \;\; t \in \mathcal{T},$$

reveals that Assumption 1 is equivalent to (9). Therefore, our results allow applied researchers to replace unconfoundedness requirement (6) for identification by the weaker condition (9) under which conditional nonsingularity is both necessary and sufficient for identification, thereby improving robustness of empirical studies in nonexperimental settings. In particular, conditional mean independence property (9) allows for any higher conditional moment of $Y(t)$ to depend on both D and X. Our identification results are thus of general interest for the vast treatment effects literature (e.g., Athey and Imbens, 2017 for a recent literature review) and complement existing results on identification of treatment effects.

Appendix A. Treatment effects modeling with non-mutually exclusive treatment regimes

For non-mutually exclusive treatment regimes, there are $\widetilde{T}-T\geq 1$ combinations of treatments in $\mathcal{T}\equiv\{1,\ldots,T\}$, denoted $\mathcal{C}(s)$ with $s\in\{T+1,\ldots,\widetilde{T}\}\equiv\widetilde{\mathcal{T}}$. For each $s\in\widetilde{\mathcal{T}}$, multiple components of D take value one jointly if treatment combination $\mathcal{C}(s)$ occurs. Define $\widetilde{D}$ a vector of dummy variables $\widetilde{D}(s)$ taking value one for $s\in\mathcal{T}$ if only treatment $s\in\mathcal{T}$ occurs, and for $s\in\widetilde{\mathcal{T}}$ if treatment combination $\mathcal{C}(s)$ occurs. A general model that gives rise to an average structural function of the form (4) is

(10)
$$Y=p(\widetilde{D})^{\mathrm{T}}\widetilde{\beta}=\widetilde{\beta}_0+\sum_{s=1}^{\widetilde{T}}\widetilde{D}(s)\widetilde{\beta}_s,\quad E(\widetilde{\beta}\mid X)=E(\widetilde{\beta}\mid\widetilde{D},X),\quad \widetilde{\beta}=(\widetilde{\beta}_0,\widetilde{\beta}_1,\ldots,\widetilde{\beta}_{\widetilde{T}})^{\mathrm{T}},$$

restricted so that heterogeneity satisfies the conditional average additivity property

$$E(\widetilde{\beta}_s\mid X)=\sum_{t=1}^{T}\mathbb{1}\{t\in\mathcal{C}(s)\}E(\widetilde{\beta}_t\mid X),\quad s\in\widetilde{\mathcal{T}}. \tag{11}$$

For $s\in\mathcal{T}$, extending the definition of $\mathcal{C}(s)$ by setting $\mathcal{C}(s)=\{s\}$ and also writing $E(\widetilde{\beta}_s\mid X)=\sum_{t=1}^{T}\mathbb{1}\{t\in\mathcal{C}(s)\}E(\widetilde{\beta}_t\mid X)$, the implied control regression function $E(Y\mid\widetilde{D},X)$ for model (10)-(11) takes the form

$$\begin{aligned}p(\widetilde{D})^{\mathrm{T}}E(\widetilde{\beta}|X)&=E(\widetilde{\beta}_0\mid X)+\sum_{s=1}^{\widetilde{T}}\widetilde{D}(s)\left[\sum_{t=1}^{T}\mathbb{1}\{t\in\mathcal{C}(s)\}E(\widetilde{\beta}_t\mid X)\right]\\&=E(\widetilde{\beta}_0\mid X)+\sum_{t=1}^{T}\left[\sum_{s=1}^{\widetilde{T}}\mathbb{1}\{t\in\mathcal{C}(s)\}\widetilde{D}(s)\right]E(\widetilde{\beta}_t\mid X)=p(D)^{\mathrm{T}}E(\beta|X),\end{aligned}$$

the control regression function for model (1) with $\beta=(\widetilde{\beta}_0,\widetilde{\beta}_1,\ldots,\widetilde{\beta}_T)^{\mathrm{T}}$ and D such that $D(t)=\sum_{s=1}^{\widetilde{T}}\mathbb{1}\{t\in\mathcal{C}(s)\}\widetilde{D}(s)$, $t\in\mathcal{T}$. Therefore, the control regression functions, and hence also the average structural functions, for model (1) with non-mutually exclusive treatment regimes and for model (10)-(11) coincide.

Appendix B. Proofs

Preliminary result.

Lemma 1. *Suppose that $E(\|\beta\|^2)<\infty$ and Assumption 1 holds. If $E\{p(D)p(D)^T\mid X\}$ is nonsingular with probability one then $q_0(X)$ is identified.*

Proof. Let $\lambda_{\min}(X)$ denote the smallest eigenvalue of $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$. Suppose that $\overline{q}(X) \neq q_0(X)$ with positive probability on a set $\widetilde{\mathcal{X}}$, and note that $\lambda_{\min}(X) > 0$ on $\mathcal{X}$ by assumption. Then

$$\begin{aligned} E\left([p(D)^{\mathrm{T}}\{\overline{q}(X) - q_0(X)\}]^2\right) &= E\left[\{\overline{q}(X) - q_0(X)\}^{\mathrm{T}} E\{p(D)p(D)^{\mathrm{T}}|X\}\{\overline{q}(X) - q_0(X)\}\right] \\ &\geq E\left\{\|\overline{q}(X) - q_0(X)\|^2 \lambda_{\min}(X)\right\} \\ &\geq E\left\{\mathbb{1}(X \in \mathcal{X} \cap \widetilde{\mathcal{X}}) \|\overline{q}(X) - q_0(X)\|^2 \lambda_{\min}(X)\right\}. \end{aligned}$$

By definition $\Pr(\widetilde{\mathcal{X}}) > 0$ and $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$ so that $\widetilde{\mathcal{X}} \cap \mathcal{X} = \widetilde{\mathcal{X}}$. Thus the fact that $\|\overline{q}(X) - q_0(X)\|^2 \lambda_{\min}(X)$ is positive on $\widetilde{\mathcal{X}} \cap \mathcal{X}$ implies

$$E\left\{\mathbb{1}(X \in \mathcal{X} \cap \widetilde{\mathcal{X}}) \|\overline{q}(X) - q_0(X)\|^2 \lambda_{\min}(X)\right\} > 0.$$

We have shown that, for $\overline{q}(X) \neq q_0(X)$ with positive probability on a set $\widetilde{\mathcal{X}}$,

$$E\left([p(D)^{\mathrm{T}}\{\overline{q}(X) - q_0(X)\}]^2\right) > 0,$$

which implies $p(D)^{\mathrm{T}} \overline{q}(X) \neq p(D)^{\mathrm{T}} q_0(X)$. Therefore, $q_0(X)$ is identified from $E(Y \mid D, X)$. □

Proof of Theorem 1. We first show that nonsingularity of $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ with probability one implies identification of $\mu(D)$. By Lemma 1, if $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is nonsingular with probability one then $q_0(X)$ is identified, and hence $E\{q_0(X)\}$ also is. By $p(D)$ being a known function, $p(D)^{\mathrm{T}}E\{q_0(X)\} = \mu(D)$ is identified.

We now establish that nonsingularity of $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ with probability one is necessary for identification of $\mu(D)$. It suffices to show that singularity of $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ with positive probability implies that $\mu(D)$ is not identified, i.e., there exists an observationally equivalent $\overline{q}(X) \neq q_0(X)$ with positive probability such that $p(D)^{\mathrm{T}}E\{\overline{q}(X)\} \neq p(D)^{\mathrm{T}}E\{q_0(X)\}$ with positive probability. By nonsingularity of $E\{p(D)p(D)^{\mathrm{T}}\}$ and linearity of $\mu(D)$, the conclusion holds if, and only if, there exists an observationally equivalent $\overline{q}(X) \neq q_0(X)$ with positive probability such that $E\{\overline{q}(X)\} \neq E\{q_0(X)\}$.

Suppose that $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is singular with positive probability and let $\Delta(X)$ be such that $E\{p(D)p(D)^{\mathrm{T}} \mid X\}\Delta(X) = 0$. We have that $\Delta(X) \neq 0$ on a set $\widetilde{\mathcal{X}}$ with $\Pr(\widetilde{\mathcal{X}}) > 0$. For $J = T + 1$, define $\widetilde{\mathcal{X}}_j = \{x \in \widetilde{\mathcal{X}} : \Delta_j(x) \neq 0\}$, $j \in \{1, \ldots, J\}$.

Then $\cup_{j=1}^{J}\widetilde{\mathcal{X}}_j = \{x \in \widetilde{\mathcal{X}} : \Delta(x) \neq 0\} = \widetilde{\mathcal{X}}$. Hence

$$0 < \Pr(\widetilde{\mathcal{X}}) = \Pr(\cup_{j=1}^{J}\widetilde{\mathcal{X}}_j) \leq \sum_{j=1}^{J} \Pr(\widetilde{\mathcal{X}}_j),$$

which implies that $\Pr(\widetilde{\mathcal{X}}_{j^*}) > 0$ for some $j^* \in \{1, \ldots, J\}$.

Set $\widetilde{\Delta}(x) = \Delta(x)$ for $x \in \widetilde{\mathcal{X}}_{j^*}$, and $\widetilde{\Delta}(x) = 0$ otherwise. By construction $\widetilde{\Delta}_{j^*}(X) \neq 0$, and letting

$$\widetilde{\widetilde{\Delta}}(X) = \text{sign}\{\widetilde{\Delta}_{j^*}(X)\}\frac{\widetilde{\Delta}(X)}{||\widetilde{\Delta}(X)||},$$

we have that $\widetilde{\widetilde{\Delta}}_{j^*}(X) > 0$ on $\widetilde{\mathcal{X}}_{j^*}$ and $||\widetilde{\widetilde{\Delta}}(X)|| = 1$, and hence $E\{||\widetilde{\widetilde{\Delta}}(X)||\} < \infty$ and $E\{\widetilde{\widetilde{\Delta}}_{j^*}(X)\} \neq 0$. Therefore $E\{\widetilde{\widetilde{\Delta}}(X)\} \neq 0$, which implies that $E\{q_0(X) + \widetilde{\widetilde{\Delta}}(X)\} \neq E\{q_0(X)\}$. The result follows. □

Proof of Theorem 2. The matrix $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is of the form

$$E\{p(D)p(D)^{\mathrm{T}} \mid X\} = \begin{bmatrix} 1 & E(D^{\mathrm{T}} \mid X) \\ E(D \mid X) & E(DD^{\mathrm{T}} \mid X) \end{bmatrix}, \tag{12}$$

and is positive definite if, and only if, the Schur complement of 1 in (12) is positive definite (Boyd and Vandenberghe, 2004, Appendix A.5.5.), i.e., if, and only if,

$$E(DD^{\mathrm{T}} \mid X) - E(D \mid X)E(D^{\mathrm{T}} \mid X) = \text{var}(D \mid X),$$

is positive definite with probability one, as claimed. □

Proof of Theorem 3. Suppose that the matrix $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is nonsingular with probability one. For mutually exclusive treatment regimes, $D \in \{0_T, \{e_t\}_{t\in\mathcal{T}}\}$ and hence $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is of the form

$$E\{p(D)p(D)^{\mathrm{T}}|X\} = \left\{p(0_T)p(0_T)^{\mathrm{T}}\right\}\times\Pr(D = 0_T|X) + \sum_{t=1}^{T}\left\{p(e_t)p(e_t)^{\mathrm{T}}\right\}\times\Pr(D = e_t|X),$$

a sum of $T + 1$ rank one $(T + 1) \times (T + 1)$ distinct matrices which is singular with positive probability if either $\Pr(D = 0_T \mid X) = 0$ or $\Pr(D = e_t \mid X) = 0$ for some $t \in \mathcal{T}$ with positive probability. For mutually exclusive treatment regimes

$$\Pr(D = e_t \mid X) = \Pr\{D(t) = 1 \mid X\}, \quad t \in \mathcal{T},$$

and hence if either $\Pr(D = 0_T \mid X) = 0$ or $\Pr\{D(t) = 1 \mid X\} = 0$ for some $t \in \mathcal{T}$ with positive probability, then $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is singular with positive

probability. Therefore, nonsingularity of $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ with probability one implies that $\Pr(D = 0_T \mid X) > 0$ and $\Pr\{D(t) = 1 \mid X\} > 0$ for each $t \in \mathcal{T}$ with probability one. Since conditional probabilities add up to unity with probability one, for mutually exclusive treatment regimes

$$\Pr(D = 0_T \mid X) + \sum_{t=1}^{T} \Pr\{D(t) = 1 \mid X\} = 1$$

with probability one, and we have shown that $\Pr\{D(t) = 1 \mid X\} > 0$ for each $t \in \mathcal{T}$ and $\Sigma_{s=1}^{T} \Pr\{D(s) = 1 \mid X\} < 1$, with probability one.

We show the converse result. Assume that $\Pr\{D(t) = 1 \mid X\} > 0$ for each $t \in \mathcal{T}$ and $\Sigma_{s=1}^{T} \Pr\{D(s) = 1 \mid X\} < 1$, with probability one. For a vector $w \in \mathbb{R}^T$, let $\mathrm{diag}(w)$ denote the $T \times T$ diagonal matrix with diagonal elements $w_1, \ldots, w_T$. For mutually exclusive treatments, the matrix $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is also of the form

$$E\{p(D)p(D)^{\mathrm{T}} \mid X\} = \begin{bmatrix} 1 & E(D^{\mathrm{T}} \mid X) \\ E(D \mid X) & \mathrm{diag}\{E(D \mid X)\} \end{bmatrix}. \tag{13}$$

The matrix $\mathrm{diag}\{E(D \mid X)\}$ has diagonal elements $E\{D(t) \mid X\} = \Pr\{D(t) = 1 \mid X\} > 0$ for each $t \in \mathcal{T}$, by assumption, and hence is positive definite and invertible.

By assumption $\Sigma_{s=1}^{T} \Pr\{D(s) = 1 \mid X\} < 1$, and hence

$$\begin{aligned} 0 < 1 - \Sigma_{s=1}^{T} \Pr\{D(s) = 1 \mid X\} &= 1 - \Sigma_{s=1}^{T} E\{D(s) \mid X\} \\ &= 1 - E(D^{\mathrm{T}} \mid X)\mathrm{diag}\{E(D \mid X)\}^{-1} E(D \mid X). \end{aligned}$$

Thus the Schur complement of $\mathrm{diag}\{E(D \mid X)\}$ in (13) is positive definite, and hence $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is positive definite (Boyd and Vandenberghe, 2004, Appendix A.5.5.). Therefore, $E\{p(D)p(D)^{\mathrm{T}} \mid X\}$ is nonsingular with probability one, as claimed. □

Proof of Theorem 4. For mutually exclusive treatments, result (i) follows from equivalence between conditional nonsingularity and common support and the argument in the main text. For non-mutually exclusive treatments, $q_u(X)$ is identified for each $u \in (0, 1)$ by an argument similar to the proof of Lemma 1, upon substituting $q_u(X)$ for $q_0(X)$. Result (ii) then follows from the argument in the main text. □

Proof of Theorem 5. The proof is similar to the proofs of Theorems 1 and 3 and hence is omitted. □

Proof of Theorem 6. For mutually exclusive treatments, by Theorem 5 conditional nonsingularity on $\mathcal{X}(s)$ is equivalent to the support of D conditional on $X = x$ being the same as the marginal support of D for almost every $x \in \mathcal{X}(s)$. Hence, the support of X conditional on D contains $\mathcal{X}(s)$ with probability one. Result (i) then follows from the argument in the main text.

For non-mutually exclusive treatments, $q_u(X)$ is identified on $\mathcal{X}(s)$ for each $u \in (0, 1)$ by an argument similar to the proof of Lemma 1, upon substituting $q_u(X)$ for $q_0(X)$, letting $\overline{q}(X) \neq q_u(X)$ on a set with positive probability $\widetilde{\mathcal{X}} \subseteq \mathcal{X}(s)$. Result (ii) then follows from the argument in the main text. □

References

Ao, W., Calonico, S., and Lee, Y. Y. (2021). Multivalued treatments and decomposition analysis: An application to the WIA program. *J. Bus. Econ. Statist.* **39**, 358–371.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: causality and policy evaluation. *J. Econ. Perspect.* **31**, 3–32.

Becker, S. O. and Egger, P. H. (2013). Endogenous product versus process innovation and a firm's propensity to export. *Emp. Econ.* **44**, 329–354.

Blundell, R., and Powell, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics*, vol. 1. Cambridge: Cambridge University Press.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breusch, T. S. (1986). Hypothesis testing in unidentified models. *Rev. Econ. Stud.* **53**, 635–651.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *J. Economet.* **155**, 138–154.

Chamberlain, G. (1984). Panel Data. In *Handbook of Econometrics*, Z. Griliches & M. D. Intriligator, eds., vol. 2. Amsterdam: Elsevier, pp. 1247–1318.

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y. and Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statist. Med.* **30**, 681–697.

Frölich, M. (2004). Programme evaluation with multiple treatments. *J. Econ. Surv.* **18**, 181–224.

Graham, B. S. and Pinto, C. C. D. X. (2018) Semiparametrically efficient estimation of the average linear regression function. *arXiv*: 1810.12511.

Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Am. Statist. Assoc.* **99**, 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706–710.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* **77**, 1481–1512.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*, 43–58. Physica, Heidelberg.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Rev. Econ. Statist.* **84**, 205–220.

Newey, W. K. and Stouli, S. (2018) Heterogenous coefficients, discrete instruments, and identification of treatment effects. *eprint arXiv:1811.09837.*

Newey, W. and Stouli, S. (2021). Control variables, discrete instruments, and identification of structural functions. *J. Economet.* **222**, 73–88.

Nian, H., Yu, C., Ding, J., Wu, H., Dupont, W.D., Brunwasser, S., Gebretsadik, T., Hartert, T.V. and Wu, P. (2019). Performance evaluation of propensity score methods for estimating average treatment effects with multi-level treatments. *J. App. Statist.* **46**, 853–873.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press, 2nd ed.

Petropoulou, M., Efthimiou, O., Rücker, G., Schwarzer, G., Furukawa, T.A., Pompoli, A., Koek, H.L., Del Giovane, C., Rodondi, N. and Mavridis, D. (2021). A review of methods for addressing components of interventions in meta-analysis. *Plos one* **16**, e0246631.

Rosenbaum, P. R. and Rubin, D. B. (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. *J. Educ. Psychol.* **66**, 688–701.

Tortú, C., Crimaldi, I., Mealli, F. and Forastiere, L. (2020). Modelling network interference with multi-valued treatments: the causal effect of immigration policy on crime rates. *arXiv*: 2003.10525.

Uysal, S. D. (2015). Doubly robust estimation of causal effects with multivalued treatments: an application to the returns to schooling. *J. App. Economet.* **30**, 763–786.

Wooldridge, J. M. (2005). Unobserved heterogeneity and the estimation of average partial effects. In *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* . Cambridge: Cambridge University Press.