Controlling for Latent Confounding with Triple Proxies

Ben Deaner *

May 2, 2022

Abstract

We apply results in Hu & Schennach (2008) to achieve nonparametric identification of causal effects using noisy proxies for unobserved confounders. We call this the 'triple proxy' approach because it requires three proxies that are jointly independent conditional on unobservables. We consider three different choices for the third proxy: it may be an outcome, a vector of treatments, or a collection of auxiliary variables. We compare to an alternative identification strategy introduced by Miao *et al.* (2018) in which causal effects are identified using two conditionally independent proxies. We refer to this as the 'double proxy' approach. We show that the conditional independence assumptions in the double and triple proxy approaches are non-nested, which suggests that either of the two identification strategies may be appropriate depending on the particular setting.

A number of recent papers consider nonparametric identification of causal effects when two noisy proxies are available for unobserved confounding factors (for example, Miao *et al.* (2018), Deaner (2021), Kallus *et al.* (2021), and Singh (2020)). We refer to the strategy in these papers as the 'double proxy' approach. If the confounders were correctly measured then one could control for them using methods like inverse propensity score re-weighting or nonparametric regression. The problem addressed in these works is thus one of non-classical measurement error in which the mismeasured variables are controls rather than treatments.

For example, suppose we wish to assess the effect of an educational intervention on a student's high school GPA. In this setting we would like to control for academic aptitude, which could confound treatments and outcomes. While we do not observe aptitude directly, we observe test scores which are noisy and possibly biased measurements of academic ability.

The identifying assumptions of the double proxy approach bear some resemblance to those in an earlier literature initiated by Hu & Schennach (2008) (hereon HS). HS consider nonparametric identification in the presence of non-classical measurement error. HS identify the joint distribution of all proxies and

*Yale University, Cowles Foundation. Email at bendeaner@gmail.com.

underlying mismeasured variables. This joint distribution is not identified by the double proxy approach.

One important distinction is that HS require the presence of three variables which are jointly independent conditional on the latent, mismeasured factors. As such we refer to causal inference using HS as the 'triple proxy' approach in the context of mismeasured controls. Compared with the double proxy approach, the triple proxy approach requires an additional vector of proxies. However, under appropriate conditions, one can use the outcome or the vector of treatments as the third proxy.

In this work we identify causal effects using the triple proxy approach. We apply results from HS taking the third proxy to be either an outcome, a vector of treatments, or a vector of auxiliary variables. In the latter two cases, the identification argument proceeds in two steps. First we apply HS to identify distributions involving the latent confounders. We then identify objects of interest from a linear integral equation that involves densities identified in the first step. A closely related two-step strategy was previously explored in the context of regression discontinuity design in a working paper Rokkanen (2015).

A key motivation for our work is to expand the settings in which one can credibly identify causal effects using mismeasured controls. Both the double proxy and triple proxy approaches restrict the causal relationships between the treatments, outcomes, proxies, and confounders. Importantly, these exclusion restrictions are non-nested. That is, there are conditions under which the double proxy approach is applicable but not the triple proxy approach, and vice versa.

HS require a condition that at least one of the noisy proxies is say, mean or median unbiased for the latent factors. If the latent factors are controls, so that we are not interested in causal effects of the factors themselves, then we can drop this assumption. However, if this condition does hold, then the triple proxy approach allows us to identify the effects of the latent factors themselves as well as their distribution, which is not possible using the double proxy approach.

Freyberger (2021) shows that if the mean or median unbiasedness condition of HS is replaced with a related monotonicity condition, one can identify objects involving quantiles of the latent variables. We adapt the strategy of Freyberger (2021) to our setting in order to identify the causal effect of shifting the latent confounders between quantiles.

The identification results in HS require some statistical completeness assumptions. These conditions are very similar to those required for double proxy analysis.

One advantage of the triple proxy approach is that it allows researchers to directly impose assumptions directly on objects that involve the latent factors. We exploit this property to partially identify causal effects under weaker exclusion restrictions when the Conditional (on the latent confounders) Average Treatment Effect (CATE) satisfies a monotonicity condition. If the CATE is constant then the identified set is a singleton.

Notation

Before we proceed, let us define some notation. If A, B, C, and D are random variables and A and B admit a joint probability density function conditional on C and D, then $f_{AB|CD}(a, b|c, d)$ is this density evaluated at $A = a$, $B = b$, $C = c$ and $D = d$. We always use upper case letters to denote random variables while the corresponding lower case letters denote values of these random variables. If A is independent of B given c we write $A \perp\!\!\!\perp B|C$. If A is jointly independent of B and C given D and E we write $A \perp\!\!\!\perp (B, C)|(D, E)$.

If Y is an outcome variable and X a treatment, then $Y(x)$ is the potential outcome from a counterfactual level x of X. Throughout we implicitly assume that $Y(X) = Y$, a condition sometimes known as 'consistency'.

'$\overset{a.s.}{=}$' denotes almost sure equality. For example, if δ is a measurable function on $\mathcal{A}$ the support of A, then $\delta(A) \overset{a.s.}{=} 0$ means the random variable $\delta(A)$ equals zero with probability one. $\operatorname{ess\,sup}_{a \in \mathcal{A}} \delta(a)$ is the essential supremum of δ, i.e., the smallest constant c so that $\delta(A) \leq c$ with probability 1. Similarly $\operatorname{ess\,inf}_{a \in \mathcal{A}} \delta(a)$ is the essential infimum: the largest almost sure lower bound.

1 A Motivating Example

Suppose we are interested in the causal effect of an educational intervention X on a student's GPA at the end of high-school Y. Whether or not a student receives the intervention is determined by the student's teachers, parents, and perhaps the student herself. These actors base their decision, at least in part, on their private assessments of the student's academic aptitude.

In this setting, academic aptitude (at the time treatment status is decided) is an unobserved confounder. It affects the decision to treat the student and it has an effect on high-school GPA, regardless of treatment. The researcher has access to some test scores that reflect academic ability, but which do not measure it perfectly.

In sum, test scores are noisy and possibly biased measurements of an unobserved confounder (academic aptitude). The need to account for the mismeasurement of ability arises in numerous empirical applications, for example in Griliches & Mason (1972), Fruehwirth *et al.* (2016), and Deaner (2021).

Identification is complicated by the fact that test scores can directly cause or be caused by treatments and outcomes. If the educational intervention affects a student's academic progress, then it presumably affects the scores on tests taken after the intervention. If a test score is used to decide some feature of the student's education other than the intervention, then it may have an affect on GPA that is not mediated by treatment. If a test is taken prior to the intervention, then it may determine eligibility for treatment.

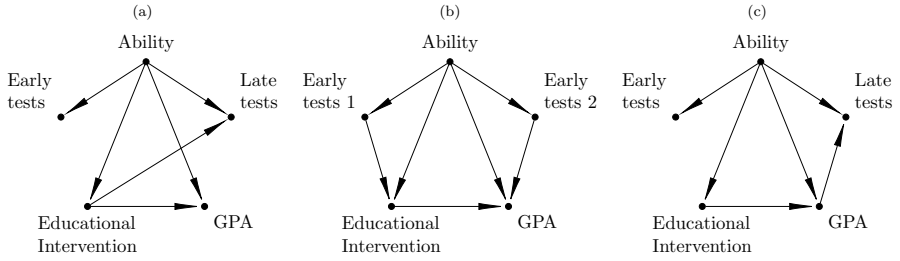
Ruling out causal relationships of this kind requires detailed institutional knowledge. In this work we show that the triple proxy approach allows for causal relationships between the proxies, treatments, and outcomes that are

incompatible with the double proxy approach, and vice versa. Thus there may be settings in which institutional knowledge is compatible with one of the two approaches but not the other.

In Figure 1 we present Directed Acyclic graphs (DAGs) that encode possible causal relationships between academic ability, the treatment (an educational intervention), the outcome (GPA at the end of high-school), and two sets of test scores.

Each DAG encodes a set of exclusion restrictions in a Non-Parametric Structural Equations Model (NPSEM) of the kind in Pearl (2009). Each NPSEM implies a set of conditional independence restrictions required for identification of causal effects using either the double or triple proxy approach.[1]

Figure 1: Test Score Proxy Graphs



The causal diagram in Figure 1.a implies the conditional independence restrictions required by both the double and triple proxy approaches. In this DAG one set of scores are from 'early tests' taken prior to the decision to treat and some are from 'late tests' taken after treatment is administered.

The DAG in Figure 1.a encodes an assumption that there is no direct effect of the tests on high-school GPA, nor on treatment. This effectively rules out the possibility that the test scores are used to determine any important aspects of a student's education. In Deaner (2018) this is justified by the fact that the test scores are only observed by researchers who have no input into the students' education.

The graph in Figure 1.a allows the educational intervention to affect the scores on post-treatment tests, this is important because if the educational intervention affects academic performance then it likely affects future test scores.

The graph in Figure 1.b is adapted from Miao *et al.* (2018). In this case one set of early tests can determine treatment. The other early tests cannot affect treatment but could impact some other aspect of the student's education and thus affect the outcome.

[1] A causal diagram can also be associated with one of various other statistical models involving counterfactual outcomes. A number of these are discussed in Robins & Richardson (2011). The results in this paper also apply when the causal diagrams are associated with any of the models discussed in Robins & Richardson (2011) including the widely used Finest Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) of Robins (1986).

Figure 1.b is compatible with the double proxy approach but not the triple proxy approach. The triple proxy approach employs Hu & Schennach (2008), which requires three proxies that are independent conditional on unobservables. Under Figure 1.b no three of the four observables are guaranteed to be jointly independent conditional on ability. Nor are there three observables that are independent conditional on ability and whichever observable is left over.

Finally, Figure 1.c is compatible with the triple proxy approach but not the double proxy approach. In this case, one set of scores are from tests taken after high-school graduation. High-school GPA could affect say, college attendance and thus later test scores. The treatment may affect these late test scores so long as this is mediated by academic progress in high-school as measured by GPA.

2 Identification in Hu and Schennach (2008)

HS prove identification of a nonparametric factor model. We restate their results below. Let W be an unobserved, possibly vector-valued latent factor with support $\mathcal{W}$. Let V, Z, and C be observable random vectors with respective supports $\mathcal{V}$, $\mathcal{Z}$, and $\mathcal{C}$. The following conditions are from Hu and Schennach (2008).

HS Assumption 1. V, Z, W, and C admit a bounded, non-zero density with respect to the product measure of the Lebesgue measure on $\mathcal{V} \times \mathcal{Z} \times \mathcal{W}$ and some dominating measure μ on $\mathcal{C}$. All marginal and conditional densities are also bounded.

HS Assumption 2. V, Z, and C are jointly independent conditional on W. Formally, $V \perp\!\!\!\perp (Z, C)|W$ and $Z \perp\!\!\!\perp C|W$.

HS Assumption 3. For any bounded function δ in $L_1(\mathcal{W})$:

$$\int_{\mathcal{W}} f_{V|W}(V|w)\delta(w)dw \overset{a.s.}{=} 0 \implies \delta(W) \overset{a.s.}{=} 0$$

and the same holds with V replaced by Z.

HS Assumption 4. For any $w_1, w_2 \in \mathcal{W}$ if $w_1 \neq w_2$ then $P\big(f_{C|W}(C|w_1) \neq f_{C|W}(C|w_2)\big) > 0$.

HS Assumption 1 ensures some bounded densities exist. HS Assumption 2 states that V, Z, and C are jointly independent conditional on W.

If the marginal densities f_W , f_V, and f_Z are bounded below away from zero over $\mathcal{W}$, $\mathcal{V}$, and $\mathcal{Z}$ respectively, then Assumption 3 is equivalent to two bounded completeness conditions. Namely, bounded completeness of W for V, and bounded completeness of W for Z. Note that Assumption 3 differs slightly from the corresponding condition in Hu & Schennach (2008), the version we use here is employed in the Handbook of Econometrics treatment of HS (see Schennach (2020)).

Statistical completeness conditions are used to identify Nonparametric Instrumental Variables (NPIV) models of the kind in Newey & Powell (2003) and Ai & Chen (2003). Thus condition 3 states that V and Z are both relevant instruments for W in the sense of NPIV.

Assumptions 4 is a relatively weak condition on the association between C and W. Hu & Schennach (2008) note that this assumption is weaker than imposing HS Assumption 3 on C. In words it states that any change in W must induce some change in the conditional distribution of C. This condition can hold even if C is a binary random variable and W is a continuous random vector.

The main identification result in HS is given below.

HS Theorem (Hu and Schennach (2008)). *Under HS Assumptions 1 and 2 the following equalities hold:*

$$f_{ZC|V}(z,c|v) = \int_{\mathcal{W}} f_{C|W}(c|w) f_{W|V}(w|v) f_{Z|W}(z|w) dw \tag{2.1}$$

$$f_{VZC}(v,z,c) = \int_{\mathcal{W}} f_{CW}(c,w) f_{V|W}(v|w) f_{Z|W}(z|w) dw \tag{2.2}$$

Moreover, under HS Assumptions 1-4, $f_{W|V}$, $f_{Z|W}$, and $f_{W|C}$ are identified from the above up to reorderings of W. Formally, suppose some other conditional densities $\tilde{f}_{W|V}$, $\tilde{f}_{Z|W}$, and $\tilde{f}_{W|C}$ satisfy Assumption 1-4 and (2.1):

$$f_{ZC|V}(z,c|v) = \int_{\mathcal{W}} \tilde{f}_{W|V}(w|v) \tilde{f}_{Z|W}(z|w) \tilde{f}_{C|W}(c|w) dw$$

Then there exists a injective function $\varphi : \mathcal{W} \to \mathcal{W}$ so that $\tilde{f}_{W|V}(w|v) = f_{\varphi(W)|V}(w|v)$, $\tilde{f}_{Z|W}(z|w) = f_{Z|\varphi(W)}(z|w)$, and $\tilde{f}_{C|W}(c|w) = f_{C|\varphi(W)}(c|w)$. And similarly for (2.2).

Theorem 1 identifies conditional densities up to reorderings of W. HS pin down a single ordering using an additional assumption given below. In the case of mismeasured control variables, causal effects are invariant to reorderings of W, and so we do not require this condition.

HS Assumption 5. There is a known functional M so that $M[f_{Z|W}(\cdot|w)] = w$, $\forall w \in \mathcal{W}$.

3 Identification with Mismeasured Controls Using Triple Proxies

To apply Hu and Schennach (2008) to identify causal effects with mismeasured controls, we suppose that two vectors of proxies V and Z are available. The third proxy C will be either the outcome Y, treatment X, or some additional observables. These choices are appropriate under different sets of exclusion restrictions.

3.1 Outcome Proxies

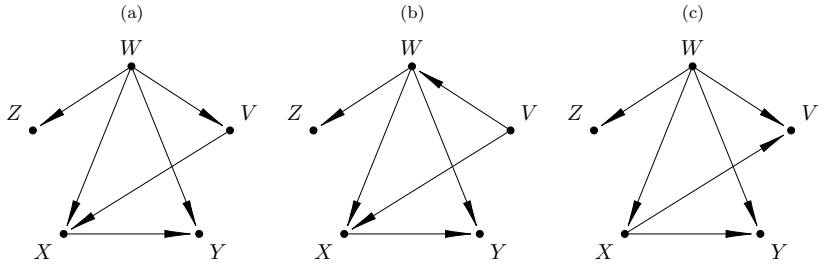
We first apply the results in Section 2 to identify $f_{Y|WX}$ and $f_{W|X}$ in a first stage. The outcome Y acts as the proxies C. We apply the results in Section 2 after conditioning on X, i.e., within each stratum of the treatment. Having identified $f_{Y|WX}$ and $f_{W|X}$ we then identify the conditional distribution of potential outcomes.

In this context we assume the existence of bounded densities akin to HS Assumption 1 in the previous section. Let $\mathcal{V}$, $\mathcal{Z}$, $\mathcal{W}$, $\mathcal{Y}$, and $\mathcal{X}$ be the supports of V, Z, W, Y, and X respectively.

Assumption 1. V, Z, W, Y, and X admit a bounded, non-zero density with respect to the product of the Lebesgue measure on $\mathcal{V} \times \mathcal{Z} \times \mathcal{W}$, some dominating measure μ_Y on $\mathcal{Y}$, and a dominating measure μ_X on $\mathcal{X}$. All marginal and conditional densities are also bounded.

Figure 2 contains three alternative causal graphs. These graphs encode exclusion restrictions on an NPSEM that imply a set of conditional independence restrictions which we use for identification.

Figure 2: Outcome Proxy Graph



The causal graphs in Figures 2.a and 2.b suggest that V is a pre-treatment variable and allow V to affect the treatment X. The graph in 2.c suggests V is a post-treatment and could be affected by treatment.

The graphs preclude X affecting Z or vice versa. This is most credible when Z is a pre-treatment variable (and thus cannot be affected by treatment), which is not used to decide treatment.

Crucially, neither the proxies Z nor V may directly affect the outcome Y. Moreover, Z must not directly affect V nor vice versa.

The graphs in Figure 2 imply a set of conditional independence restrictions given in Proposition 1. The proposition can be verified straight-forwardly using the tools in Pearl (2009). Our identification results directly assume the conditional independence restrictions in the conclusion of Proposition 1. Thus the graphs in Figure 2 can be understood to represent primitive conditions for these conditional independence restrictions.

Proposition 1. *The NPSEMs associated with the causal graphs in Figure 2 all imply the following conditional independence restrictions:*

i. $Y \perp\!\!\!\perp (V, Z)|(W, X)$*, ii.* $V \perp\!\!\!\perp Z|(W, X)$*, iii.* $Z \perp\!\!\!\perp X|W$*, and iv.* $Y(x) \perp\!\!\!\perp (X, V)|W$

The conditional independence restrictions in Proposition 1 are stronger than those required for the double proxy approach. In particular, the double proxy approach requires conditions ii., iii., and iv. but not condition i.

In this setting we use the Assumption 2 below in place of HS Assumption 3.

Assumption 2. For each $x \in \mathcal{X}$ and any bounded function δ in $L_1(\mathcal{W})$:

$$\int_{\mathcal{W}} f_{V|WX}(V|w, x)\delta(w)dw \overset{a.s.}{=} 0, \implies \delta(W) \overset{a.s.}{=} 0$$

and the same holds with V replaced by Z.[2]

Assumption 2 differs from HS Assumption 3 in that it must hold within each stratum of the treatment X. If the conditional densities $f_{W|X}$, $f_{V|X}$, and $f_{Z|X}$ are all bounded below away from zero and have bounded supports, then Assumption 2 is equivalent to the completeness conditions in Deaner (2021).

Finally, Assumption 3 below plays the role of HS Assumption 4. Note that this condition allows for the possibility that the outcome Y is binary, even if W is a continuous random vector.

Assumption 3. For all $x \in \mathcal{X}$ and any $w_1, w_2 \in \mathcal{W}$, if $w_1 \neq w_2$ then:

$$P\big(f_{Y|WX}(Y|w_1, x) \neq f_{Y|WX}(Y|w_2, x)\big) > 0$$

We now identify causal effects under the conditions above.

Theorem 1 (Outcome Proxies). *Suppose Assumptions 1-3 and conclusions i., ii., and iii. of Proposition 1 and hold. Then* $f_{Y|WX}$*,* $f_{Z|W}$*, and* $f_{W|VX}$ *(and thus* $f_{W|X}$*) are identified up to reorderings of* W *from the equation below:*

$$f_{YZ|VX}(y, z|v, x) = \int_{\mathcal{W}} f_{Y|WX}(y|w, x)f_{Z|W}(z|w, x)f_{W|VX}(w|v, x)dw$$

The marginal and conditional distributions of potential outcomes are then identified from the following equalities, which are invariant to reorderings of W*:*[3]

$$f_{Y(x_1)|X}(y|x_2) = \int_{\mathcal{W}} f_{Y|WX}(y|w, x_1)f_{W|X}(w|x_2)dw \tag{3.1}$$

$$f_{Y(x_1)}(y) = \int_{\mathcal{X}} f_{Y(x_1)}(y|x)f_X(x)d\mu_X(x) \tag{3.2}$$

[2]Strictly speaking the condition only needs to hold for μ_X-almost all x rather than all $x \in \mathcal{X}$. Similarly for Assumption 3.

[3]Strictly speaking, we identify $f_{Y(x_1)|X}(y|x_2)$ and $f_{Y(x_1)}(y)$ for μ_Y-almost all y and μ_X-almost all x_1 and x_2.

Theorem 1 first identifies $f_{Y|WX}$, $f_{Z|W}$, and $f_{W|VX}$ up to reorderings of W. The conditional and marginal distribution of potential outcomes is then given in terms of these objects. Note if the binary treatment case, identification of $f_{Y(x)|X}$ for $x = 0, 1$ immediately implies identification of the average treatment effects and the average effect of treatment on the treated.

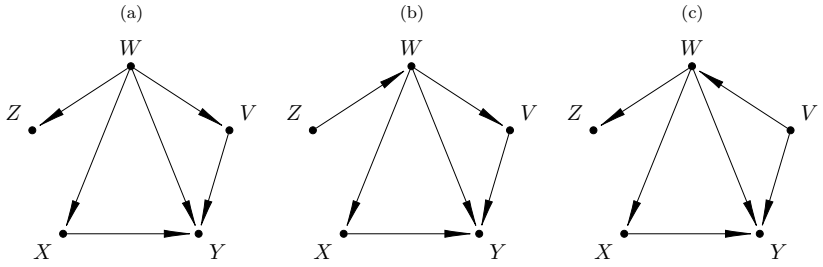
Without Conclusion iii. of Proposition 1, we could still apply Hu & Schennach (2008) to identify $f_{Y|WX}(\cdot|\cdot, x)$, $f_{Z|WX}(\cdot|\cdot, x)$, and $f_{W|VX}(\cdot|\cdot, x)$ up to reorderings of W for each x in the support of X. However, the reorderings of W could differ between the values of x. We revisit this possibility in Section 5 and show that partial identification (and possibly point identification) can be achieved without condition iii. under a monotonicity assumption.

3.2 Treatment Proxies

We now consider the case in which the third proxy C, is the vector of treatments X. In this case identification proceeds in two stages. In a first stage we use results from HS to identify conditional distributions involving W. In a second step, the distribution of potential outcomes is identified via a linear integral equation. This two-step approach is similar to one employed in Rokkanen (2015).

In this case, the causal diagrams below are sufficient for the conditional independence restrictions under which we establish identification.

Figure 3: Treatment Proxy Graphs



The diagrams in Figure 3 suggest that V is determined prior to the outcome, and the graphs allow V to directly affect the outcome. However, V and Z cannot directly affect, or be directly affected by, the treatment X. This contrasts with the double proxy case, which allows one of the two proxies to be directly causally connected to the treatment.

Proposition 2. *The NPSEMs associated with the causal graphs in Figure 3 imply the following conditional independence restrictions:*

i. $V \perp\!\!\!\perp (X, Z)|W$, *ii.* $X \perp\!\!\!\perp Z|W$, *iii.* $Y \perp\!\!\!\perp Z|(W, X)$, *and iv.* $Y(x) \perp\!\!\!\perp (X, Z)|W$

Conditions i., and iv. in Proposition 2 are those required for the double proxy approach. However, the double proxy approach does not require condition ii.

Strictly speaking, condition iii. is not required for the double proxy approach. However, if condition iv. is strengthened slightly to $Y(\cdot) \perp\!\!\!\perp (Z, X)|W$ this implies condition iii.

Assumption 4 replaces HS Assumption 4. Note that Assumption 4 may hold even if the treatment X is binary.

Assumption 4. For any $w_1, w_2 \in \mathcal{W}$ if $w_1 \neq w_2$ then $P\big(f_{X|W}(X|w_1) \neq f_{X|W}(X|w_2)\big) > 0$.

Theorem 2 (Treatment Proxies). *Suppose conclusions i., and ii. of Proposition 2, HS Assumption 3, and Assumptions 1 and 4 hold. Then f_{XW}, $f_{V|W}$, and $f_{Z|W}$ (and thus $f_{X|W}$) are identified up to reorderings of W from the equation below:*

$$f_{XVZ}(x, v, z) = \int_{\mathcal{W}} f_{XW}(x, w) f_{V|W}(v|w) f_{Z|W}(z|w) dw$$

In addition, suppose conclusion iii. of Proposition 2 holds. $f_{Y|WX}$ is then identified up to reorderings of W from the linear integral equation below:

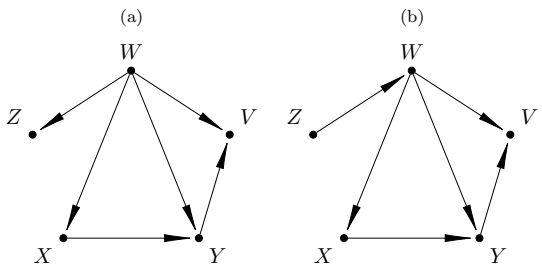
$$f_{YZX}(y, z, x) = \int_{\mathcal{W}} f_{Y|WX}(y|w, x) f_{XW}(x, w) f_{Z|W}(z|w) dw$$

If conclusion iv. of Proposition 2 holds, the conditional and marginal distributions of potential outcomes are then given by (3.1) and (3.2) respectively.

3.3 Outcome-Conditional Treatment Proxies

We again consider the case in which the third proxy C, is the vector of treatments X. However, we apply Hu & Schennach (2008) within each stratum of the outcome. This allows for the possibility that the outcome directly affects one of the proxies V, which is generally incompatible with the double proxy approach.

Figure 4: Treatment Proxy Graphs



The diagrams in Figure 4 differ from those in Figure 3 in that V is a post-outcome variable and can be impacted directly by the outcome. Note that V is a post treatment variable but X must not affect V directly.

Recall the test score example with the outcome Y measuring GPA in the final year of high-school. Suppose the tests in V taken a year after high-school graduation. GPA may affect college attendance which could in turn impact scores on the college-age tests V. The educational intervention X can influence post-high school test scores, so long as the effect of the intervention is mediated by academic achievement over high school as measured by final GPA.

Proposition 3. *The NPSEMs associated with the causal graphs in Figure 4 imply the following conditional independence restrictions:*

i. $V \perp\!\!\!\perp (X, Z)|(W, Y)$*, ii.* $X \perp\!\!\!\perp Z|(W, Y)$*, iii.* $Y \perp\!\!\!\perp Z|W$*, and iv.,* $Y(x) \perp\!\!\!\perp X|W$*.*

The conditions in Proposition 3 are insufficient for the double proxy approach. The double proxy approach requires that $V \perp\!\!\!\perp (X, Z)|W$ which generally rules out V having a direct causal effect on Y (unless we were to assume X has no causal effect on Y which defeats the purpose of our analysis). Conversely, the double proxy approach does not require any idependence between X and Z, conditional or otherwise.

In this setting we need HS Assumption 3 to apply within each stratum of the outcome.

Assumption 5. For each $y \in \mathcal{Y}$ and any bounded function δ in $L_1(\mathcal{W})$:

$$\int_{\mathcal{W}} f_{V|WY}(V|w, y)\delta(w)dw \overset{a.s.}{=} 0, \implies \delta(W) \overset{a.s.}{=} 0$$

and the same holds with V replaced by Z.

Finally, we need HS Assumption 4 to hold for the treatment proxy within each stratum of Y.

Assumption 6. For all $y \in \mathcal{Y}$ and any $w_1, w_2 \in \mathcal{W}$, if $w_1 \neq w_2$ then:

$$P\big(f_{X|WY}(X|w_1, y) \neq f_{X|WY}(X|w_2, y)\big) > 0$$

Theorem 3 (Conditional Treatment Proxies). *Suppose conclusion i., ii., and iii, of Proposition 3 and Assumptions 1, 5, and 6 hold. Then* $f_{X|WY}$*,* $f_{Z|W}$*, and* $f_{W|VY}$ *are identified up to reorderings of* W *from the equation below:*

$$f_{XZ|VY}(x, z|v, y) = \int_{\mathcal{W}} f_{X|WY}(x|w, y)f_{Z|W}(z|w)f_{W|VY}(w|v, y)dw$$

$f_{W|X}$ *can be written in terms of* $f_{X|WY}$*,* $f_{W|VY}$ *and the joint distribution of observables and is thus also identified up to reorderings of* W*.*[4] $f_{Y|WX}$ *is then identified up to reorderings of* W *from the linear integral equation below:*

$$f_{YZ|X}(y, z|x) = \int_{\mathcal{W}} f_{Y|WX}(y|w, x)f_{Z|W}(z|w)f_{W|X}(w|x)dw \tag{3.3}$$

[4]More precisely, $f_{W|X}$ is given by the following equation:

$$f_{W|X}(w|x) = \int_{\mathcal{Y}}\int_{\mathcal{V}} f_{X|WY}(x|w, y)f_{W|VY}(w|v, y)\frac{f_{VY}(v, y)}{f_X(x)}dvdy$$

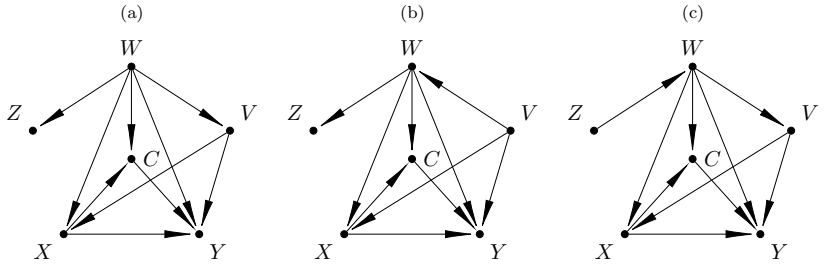
In addition, suppose conclusion iv. of Proposition 3 holds. The conditional and marginal distributions of potential outcomes are then given by (3.1) and (3.2) respectively.

3.4 Auxiliary Proxies

Finally, we consider the case in which the third proxy C is a vector of auxiliary variable (as opposed to X or Y). In this case we apply the results from Section 2 within each stratum of the treatments.

Assumption 7. V, Z, W, Y, X, and C admit a bounded, non-zero density with respect to the product of the Lebesgue measure on $\mathcal{V} \times \mathcal{Z} \times \mathcal{W}$, some dominating measure μ_Y on $\mathcal{Y}$, a dominating measure μ_X on $\mathcal{X}$, and a dominating measure μ_C on $\mathcal{C}$. All marginal and conditional densities are also bounded.

Figure 5: Auxiliary Proxy Graphs



The causal graphs in Figure 5 provide the key exclusion restrictions in this setting. The strongest restrictions in Figure 5 are on Z. Z cannot directly cause or be caused by treatment or outcome.

Suppose C is a post-treatment proxy and V is a pre-treatment proxy, and both are determined prior to the outcome Y. In addition, let us assume that W is determined prior to all the observables, with the possible exception of V. Then V and C can be caused by all variables determined prior to them other than Z and can determine all variables determined after them other than Z.

The exclusion restrictions in Figure 5 are not sufficient for the independence restrictions of the double proxy approach. In the double proxy approach each of the two proxies must be independent of either the treatment or outcome. This effectively rules out any proxies being directly causally related to both X and Y. Figure 5 allows C and V to be causally related to both X and Y and so neither can act as proxy in the double proxy case.

Proposition 4. The NPSEMs associated with the causal graphs in Figure 5 imply the following conditional independence restrictions:

i. $C \perp\!\!\!\perp (V, Z)|(W, X)$, ii. $V \perp\!\!\!\perp Z|(W, X)$, iii. $X \perp\!\!\!\perp Z|W$, iv. $Y \perp\!\!\!\perp Z|(W, V, X)$, and v. $Y(x) \perp\!\!\!\perp X|(W, V)$.

In this setting Assumption 8 replaces HS Assumption 4.

Assumption 8. For all $x \in \mathcal{X}$ and any $w_1, w_2 \in \mathcal{W}$, if $w_1 \neq w_2$ then:

$$P\big(f_{C|WX}(C|w_1, x) \neq f_{C|WX}(C|w_2, x)\big) > 0$$

Theorem 4 (Auxiliary Proxies). *Suppose conclusions i., ii., and iii. of Proposition 4 and Assumptions 2, 7, and 8 hold. Then $f_{C|WX}$, $f_{W|VX}$, and $f_{Z|W}$ are identified up to reorderings of W from the equation below:*

$$f_{CZ|VX}(c, z|v, x) = \int_{\mathcal{W}} f_{C|WX}(c|w, x) f_{W|VX}(w|v, x) f_{Z|W}(z|w) dw$$

In addition, if conclusion iv. of Proposition 4 holds, $f_{Y|WVX}$ is identified (up to reorderings of W) from the linear integral equation below:

$$f_{YZ|VX}(y, z|v, x) = \int_{\mathcal{W}} f_{Y|WVX}(y|w, v, x) f_{Z|W}(z|w) f_{W|VX}(w|v, x) dw$$

Under conclusion v. of Proposition 4, The conditional and marginal distributions of potential outcomes are then given by:

$$f_{Y(x_1)}(y|x_2) = \int_{\mathcal{W}} \int_{\mathcal{V}} f_{Y|WVX}(y|w, v, x_1) f_{W|VX}(w|v, x_2) f_{V|X}(v|x_2) dv dw$$

$$f_{Y(x_1)}(y) = \int_{\mathcal{X}} f_{Y(x_1)}(y|x) f_X(x) d\mu_X(x)$$

The identification result in Theorem 3 combines aspects of Theorems 1 and 2. As in Theorem 1 (outcome proxies) the application of HS is carried out while conditioning on the treatments X. Like Theorem 2 (treatment proxies) identification involves the solution to a second integral equation which contains objects obtained during the initial HS step.

4 Confounder Effects

Our analysis is premised upon the idea that the objects of interest are effects of the treatment X, and not the effects of the vector of latent confounders W. However, effects of W may be of some interest. Under some additional conditions, causal effects of W are identified as a byproduct of the results in the previous section.

Let $Y(x, w)$ denote the potential outcome under a counterfactual in which X and W are respectively set to values x and w. Proposition 5 provides conditional independence restrictions involving $Y(x, w)$ which follow from the causal graphs in the previous section.

Proposition 5. *The NPSEMs associated with all of the causal graphs in Figures 3, 4.a, 4.b, and 5 imply i. $Y(x, w) \perp\!\!\!\perp (X, W)$. The NPSEMs associated with the graphs in Figure 4.c and Figure 6 imply ii. $Y(x, w) \perp\!\!\!\perp (X, W)|V$.*

In Theorems 1, 2, and 3, $f_{Y|WX}$ is identified up to reorderings of W. In Theorem 4, $f_{Y|WVX}$ is identified up to reorderings of W. When the conclusions of Proposition 5 holds, these objects then yield identification of the distribution of $Y(x,w)$ up to reorderings of W.

Lemma 1. *Suppose the conclusion i. of Proposition 5 holds. Then:*

$$f_{Y(x,w)}(y) = f_{Y|WX}(y|w,x)$$

If conclusion ii. of Proposition 5 holds then:

$$f_{Y(x,w)}(y) = \int_{\mathcal{V}} f_{Y|WVX}(y|w,v,x) f_V(v) dv$$

In order to pin down a particular ordering of W we can use HS Assumption 5 which we repeat below.

HS Assumption 5. There is a known functional M so that $M[f_{Z|W}(\cdot|w)] = w$, $\forall w \in \mathcal{W}$.

Assumption 10 allows us to fix the correct ordering of W. If the functional M returns the mean of the distribution in its argument then the assumption states that Z is mean-unbiased for W. If M returns the median, then the assumption states that Z is median-unbiased for W. It is implicit in the assumption that the dimensions of W and Z are the same.

Freyberger (2021) shows that one can replace HS Assumption 5 with a related monotonicity condition and still identify the effect of changing W from one of its quantiles to another. For example, suppose W is a scalar that represents a student's skill at mathematics. Under a weaker condition than HS Assumption 5 we can identify the causal effect of increasing math skill from the 25-th to the 50-th percentile.

We apply the general approach of Freyberger (2021) in our setting. Assumption 11 below is similar to the monotonicity condition in Freyberger (2021). We assume monotonicity of a known functional $M[f_{Z|W}(\cdot|w)]$, Freyberger (2021) instead assumes that Z can be written as a strictly monotone function of W and some independent noise.

Assumption 9 (Freyberger). There is a known functional M so that for some (unknown) function ϕ, $M[f_{Z|W}(\cdot|w)] = \phi(w)$, $\forall w \in \mathcal{W}$. $\phi(w)$ has the same length as W, each coordinate of $\phi(w)$ depends only on the corresponding coordinate of w, and each coordinate is strictly increasing in the corresponding coordinate of w.

In order to state the main result in this section we need to introduce additional notation. Let τ be a vector of the same length as W whose k-th component τ_k, is in $[0,1]$ for each k. Let W_k be the k-th component of W. Then $Q_W(\tau)$ is a vector of the same length as W whose k-th components is the τ_k-th quantile of W_k.

Theorem 5 (Latent Confounder Effects). *Suppose the conditions of any of Theorems 1, 2, 3 and conclusion i. of Proposition 5 holds, or the conditions of Theorem 4 and conclusion ii of Proposition 5 hold. Then $f_{Y(x,w)}$, is identified up to reorderings of W.*

If HS Assumption 5 or Assumption 9 holds then $f_{Y(x,Q_W(\tau))}$ is identified. If Assumption 9 holds then $f_{Y(x,w)}$ is identified.[5]

The first part of Theorem 5 follows straight-forwardly from Lemma 1 and Theorems 1, 2, 3, and 4. These results identify $f_{Y(x,w)}$, $f_{Z|W}$, and f_W up to reorderings of W.

In the second part of the Theorem, HS Assumptions 5 and Assumption 9 are used to pin orderings of W (up to strictly increasing transformations in the latter case). We provide some detail below.

Let $\tilde{f}_{Y(x,w)}$, $\tilde{f}_{Z|W}$, and $\tilde{f}_W$ differ from $f_{Y(x,w)}$, $f_{Z|W}$, and f_W in the ordering of W. Define $\alpha(w) = M[\tilde{f}_{Z|W}(\cdot|w)]$ and let $Q_{\alpha(\tilde{W})}(\tau)$ be the quantiles of $\alpha(\tilde{W})$ where $\tilde{W}$ is a random variable with density $\tilde{f}_W$. Under Assumption 9, $f_{Y(x,Q_W(\tau))}$ is identified from:

$$\tilde{f}_{Y(x,\alpha^{-1}(Q_{\alpha(\tilde{W})}(\tau)))}(y) = f_{Y(x,Q_W(\tau))}(y)$$

Under HS Assumption 5 the true distribution of potential outcomes is identified from the following equation:

$$f_{Y(x,w)}(y) = \tilde{f}_{Y(x,\alpha^{-1}(w))}(y)$$

5 Weakening the Exclusion Restrictions Under CATE Monotonicity

In this section we note that under a monotonicity assumption on the Conditional (on W) Average Treatment Effect (CATE) we can partially identify conditional and unconditional average treatment effects under weaker exclusion restrictions than those in Section 3.

More precisely, we are able to weaken the exclusion restrictions in the outcome proxy and auxiliary proxy cases explored in Sections 3.1 and 3.4. When the CATE is constant we achieve point identification.

We assume throughout that treatment is binary with $X = 1$ indicating treatment and $X = 0$ no treatment. However, the results extend straightforwardly to more general discrete treatments.

For some intuition, recall that the identification results in Sections 3.1 and 3.4 require that Z not cause or be caused by X. The need for this exclusion restriction arises because we apply Hu & Schennach (2008) within each stratum of X and identify objects up to reorderings of W. The exclusion restriction helps to ensure that the reorderings do not differ between strata of the treatment.

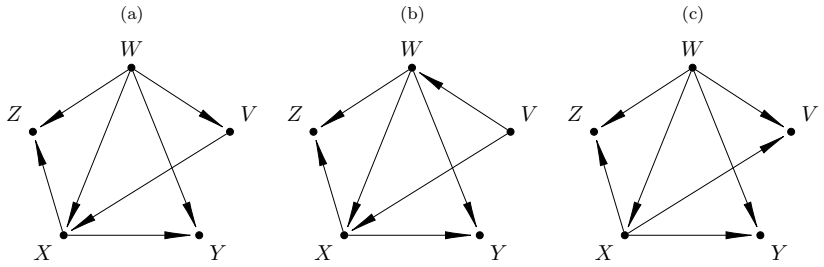
[5] Strictly speaking $f_{Y(x,Q_W(\tau))}(y)$ and $f_{Y(x,w)}(y)$ are identified for μ_Y-almost all y, μ_X-almost all x, and almost all τ and w.

Monotonicity of the CATE allows us to compare conditional average potential outcomes between treated and untreated individuals even when the ordering of W varies with treatment status.

5.1 Outcome Proxies with Monotone CATE

We apply Hu & Schennach (2008) with $C = Y$ as in Section 3.1. We weaken the exclusion restrictions in Figure 2 to those in Figure 6.

Figure 6: Outcome Proxy Graph



The graphs in Figure 6 weaken those in Figure 2 by allowing treatment X to directly impact Z which is a vector of post-treatment proxies. Consider Figure 6.a, in the test score case, V is a vector of pre-treatment test scores that can directly determine treatment and Z is a vector of post-treatment scores which can be directly affected by treatment.

The restrictions in Figure 6 are not sufficiently strong for the double proxy approach. Figure 6 allows treatment to be directly causally related to both the proxies Z and V, which is incompatible with the double proxy approach.

Proposition 6. *The NPSEMs associated with the causal graphs in Figure 6 all imply the following conditional independence restrictions:*
i. $Y \perp\!\!\!\perp (V, Z)|(W, X)$*, ii.* $V \perp\!\!\!\perp Z|(W, X)$*, and iii.* $Y(x) \perp\!\!\!\perp (X, V)|W$

The conclusions of Proposition 6 are weaker than those of Proposition 1. In particular, we drop conclusion iii. of Proposition 1 (independence of X and Z conditional on W). In contrast to the double proxy approach, V is not required to be independent of X conditional on any of the other variables.

In order to partially identify conditional average treatment effects we require a monotonicity assumption given below.

Assumption 10 (Monotone CATE). There is a constant $c < \infty$ so that $|E[Y(0)|W]|$ is almost surely bounded by c. Moreover, for any $w_1, w_2 \in \mathcal{W}$, if $E[Y(0)|W = w_2] \geq E[Y(0)|W = w_1]$ then:

$$E[Y(1) - Y(0)|W = w_2] \geq E[Y(1) - Y(0)|W = w_1]$$

Assumption 10 states that if the average untreated outcome is larger in one stratum of W than another, then the average treatment effect in that stratum is also larger. If a large value of Y indicates a more favorable outcome, then loosely speaking, those who do better without treatment tend to benefit more from treatment.

Note that the monotonicity need not be strict. Assumption 10 allows for the possibility that $E[Y(1) - Y(0)|W = w]$ is constant for all w.

Theorem 6 (monotone CATE with outcome proxies). *Suppose the conclusions of Proposition 6 holds and Assumption 1, 2, and 3, hold. Then for $x = 1, 2$, $f_{Y(x)|W}$ is identified up to reorderings of W which may depend on x.*

Thus we identify functions $\tilde{f}_{Y(1)|W}$ and $\tilde{f}_{Y(0)|W}$ which differ from $f_{Y(1)|W}$ and $f_{Y(0)|W}$ in the orderings of W. Define $\bar{s}$ and $\underline{s}$ as follows:

$$\bar{s} = \operatorname*{ess\,sup}_{w \in \mathcal{W}} \int_{\mathcal{Y}} y \tilde{f}_{Y(1)|W}(y|w)dy - \operatorname*{ess\,sup}_{w \in \mathcal{W}} \int_{\mathcal{Y}} y \tilde{f}_{Y(0)|W}(y|w)dy$$

$$\underline{s} = \operatorname*{ess\,inf}_{w \in \mathcal{W}} \int_{\mathcal{Y}} y \tilde{f}_{Y(1)|W}(y|w)dy - \operatorname*{ess\,inf}_{w \in \mathcal{W}} \int_{\mathcal{Y}} y \tilde{f}_{Y(0)|W}(y|w)dy$$

Then $\bar{s}$ and $\underline{s}$ are respectively the essential supremum and infimum of $E[Y(1) - Y(0)|W = w]$ and for $x = 0, 1$:

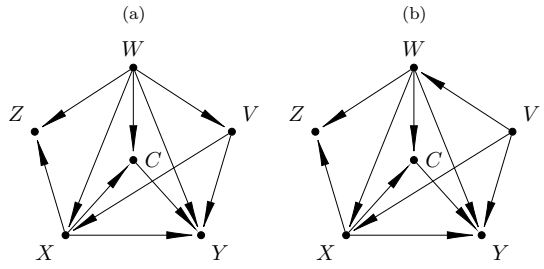
$$E[Y(1) - Y(0)|X = x] \in [\underline{s}, \bar{s}]$$

Theorem 6 partially identifies the CATE, the average effect of treatment on the treated, and the average effect of treatment on the untreated. If the CATE is constant then $\underline{s} = \bar{s}$ and so the identified set is a singleton and the effects are point identified.

5.2 Auxiliary Proxies with Monotone CATE

Finally we revisit the case examined in Section 3.4 in which C is a vector of additional variables. We weaken the exclusion restrictions in Figure 5 to those in Figure 7.

Figure 7: Auxiliary Proxy Graphs



The causal graphs in Figure 7 differ from those in Figures 6.a and 6.b in that they allow Z to be a post-treatment variable that may be directly affected treatment. Note that Z still cannot directly affect the outcome.

Note that we do not include a relaxed version on Figure 6.c. If we relaxed 7.c to allow X to affect Z the resulting graph would be cyclic.

The graphs in Figure 7 imply conditional independence restrictions given in Proposition 7. These conditions weaken those in Proposition 4. In particular we drop condition iv. of the proposition and leave the remaining conditions unchanged.

Proposition 7. *The NPSEMs associated with the causal graphs in Figure 7 imply the following conditional independence restrictions:*

i. $C \perp\!\!\!\perp (V, Z)|(W, X)$*, ii.* $V \perp\!\!\!\perp Z|(W, X)$*, iii.* $Y \perp\!\!\!\perp Z|(W, V, X)$*, and iv.* $Y(x) \perp\!\!\!\perp X|(W, V)$*.*

In this setting we adapt Assumption 10 to apply within each stratum of V. We state this condition as Assumption 11 below.

Assumption 11 (V-conditional Monotone CATE). For almost all $v \in \mathcal{V}$ there is a constant $c < \infty$ so that $|E[Y(0)|W, V = v]| < c$ with probability 1. Moreover, for any $w_1, w_2 \in \mathcal{W}$, if $E[Y(0)|W = w_2, V = v] \geq E[Y(0)|W = w_1, V = v]$ then:

$$E[Y(1) - Y(0)|W = w_2, V = v] \geq E[Y(1) - Y(0)|W = w_1, V = v]$$

As in the outcome proxy case, monotonicity of the CATE allows us to partially identify causal effects. the identified set reduces to a point when for almost all $v \in \mathcal{V}$, $E[Y(1) - Y(0)|W = w, V = v]$ does not depend on w.

Theorem 7 (monotone CATE with auxiliary proxies). *Suppose conclusions of Proposition 7 and Assumptions 2, 7, and 8 hold. Then for* $x = 1, 2$*,* $f_{Y(x)|WV}$ *is identified up to reorderings of* W *which may depend on* x*.*

Thus we identify functions $\tilde{f}_{Y(1)|WV}$ *and* $\tilde{f}_{Y(0)|WV}$ *which differ from* $f_{Y(1)|WV}$ *and* $f_{Y(0)|WV}$ *in the orderings of* W*. Define* $\bar{s}(v)$ *and* $\underline{s}(v)$ *as follows:*

$$\bar{s}(v) = \operatorname*{ess\,sup}_{w \in \mathcal{W}} \int_{\mathcal{Y}} y \tilde{f}_{Y(1)|WV}(y|w, v)dy - \operatorname*{ess\,sup}_{w \in \mathcal{W}} \int_{\mathcal{Y}} y \tilde{f}_{Y(0)|WV}(y|w, v)dy$$

$$\underline{s}(v) = \operatorname*{ess\,inf}_{w \in \mathcal{W}} \int_{\mathcal{Y}} y \tilde{f}_{Y(1)|WV}(y|w, v)dy - \operatorname*{ess\,inf}_{w \in \mathcal{W}} \int_{\mathcal{Y}} y \tilde{f}_{Y(0)|WV}(y|w, v)dy$$

Then for almost all $v \in \mathcal{V}$*,* $\bar{s}(v)$ *and* $\underline{s}(v)$ *are respectively the essential supremum and infimum of* $E[Y(1) - Y(0)|W = w, V = v]$ *and for* $x = 0, 1$*:*

$$E[Y(1) - Y(0)|X = x] \in \left[E[\underline{s}(V)|X = x], E[\bar{s}(V)|X = x]\right]$$

6 Conclusion and Further Comparison with Double Proxies

In this work we establish identification of causal objects using the 'triple proxy' approach. We show that there are sets of exclusion restrictions under which

we can establish identification (or partial identification) using the triple proxy approach, but not using double proxies. Conversely, there are exclusion restrictions that support the double proxy but not the triple proxy approach. This suggests that the double and triple proxy approaches could each be appropriate in different empirical settings.

In some cases the exclusion restrictions may allow for both approaches, see the discussion that corresponds to Figure 2 in Section 1. In this case a comparison of the merits of the two strategies is more nuanced.

On the one hand, the triple proxy approach has the advantage that under some additional conditions, it identifies causal effects of the latent variables themselves (see Section 4). In addition, the triple proxy approach enables researchers to impose a priori restrictions on densities involving the latent confounders. For example, one could constrain the solutions to the Hu & Schennach (2008) integral equation to be smooth or log-concave. The triple proxy approach also avoids the need to assume unintuitive regularity conditions that guarantee the existence of solutions to integral equations in the double proxy approach.

However, a major advantage of the double proxy over the triple proxy approach is that it motivates (relatively) simple non-parametric estimation. For example, Deaner (2021) suggests an estimator using double proxies that is similar to sieve two-stage least squares. Nonparametric estimation using the triple proxy approach is likely at least as complicated as, and possibly more complicated than, the estimator in Hu & Schennach (2008). As such, the nonparametric identification results presented here may act as motivation for a simpler parametric estimation strategy or estimation of a discretized version of the model. Nonetheless, we intend to explore nonparametric estimation using triple proxies in future work.

References

Ai, Chunrong, & Chen, Xiaohong. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**(6), 1795–1843.

Deaner, Ben. 2018. Nonparametric Estimation and Identification in Non-Separable Models Using Panel Data. Sept.

Deaner, Ben. 2021. Proxy Controls and Panel Data. Jan.

Freyberger, Joachim. 2021. Normalizations and misspecification in skill formation models. Apr.

Fruehwirth, Jane Cooley, Navarro, Salvador, & Takahashi, Yuya. 2016. How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects. *Journal of Labor Economics*, **34**, 979–1021.

Griliches, Zvi, & Mason, William M. 1972. Education, Income, and Ability. *Journal of Political Economy*, **80**, S74–S103.

Hu, Yingyao, & Schennach, Susanne M. 2008. Instrumental Variable Treatment of Nonclassical Measurement Error Models. *Econometrica*, **76**, 195–216.

Kallus, Nathan, Mao, Xiaojie, & Uehara, Masatoshi. 2021. Causal Inference Under Unmeasured Confounding With Negative Controls: A Minimax Learning Approach. Mar.

Miao, Wang, Geng, Zhi, & Tchetgen, Eric J. Tchetgen. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, **105**, 987–993.

Newey, Whitney K., & Powell, James L. 2003. Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, **71**, 1565–1578.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference (Second Edition)*. Cambridge University Press.

Robins, James. 1986. A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–1512.

Robins, James M., & Richardson, Thomas S. 2011. *Alternative Graphical Causal Models and the Identification of Direct Effects*. Oxford University Press. Chap. 6.

Rokkanen, Miikka AT. 2015. *Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design*. Working paper.

Schennach, Susanne M. 2020. *Mismeasured and unobserved variables, Handbook of Econonometrics*. Elsevier. Chap. 6, pages 487–565.

Singh, Rahul. 2020. Kernel Methods for Unobserved Confounding: Negative Controls, Proxies, and Instruments. Dec.

A Proofs

Proof of Theorem 1. Under conditions i. and ii. of Proposition 1 and Assumptions 1, 2, and 3, we can apply HS Theorem 1 conditional on $X = x_1$ with $C = Y$. This yields identification of $f_{Y|WX}(\cdot|\cdot, x_1)$, $f_{Z|WX}(\cdot|\cdot, x_1)$, and $f_{W|VX}(\cdot|\cdot, x_1)$ up to reorderings of W from the equation below:

$$f_{YZ|VX}(y, z|v, x_1) = \int_{\mathcal{W}} f_{Y|WX}(y|w, x_1) f_{Z|W}(z|w) f_{W|VX}(w|v, x_1) dw \quad \text{(A.1)}$$

In the above we have imposed that $f_{Z|W}(z|w) = f_{Z|WX}(z|w, x_1)$ which follows from conclusion iii. of Proposition 1. Now, taking (A.1) with x_2 in place of x_1, and integrating over $y \in \mathcal{Y}$ we get:

$$f_{Z|VX}(z|v, x_2) = \int_{\mathcal{W}} f_{Z|W}(z|w) f_{W|VX}(w|v, x_2) dw$$

We have already identified $f_{Z|W}$, and $f_{Z|VX}$ only involves observables, so the only unknown in the above is $f_{W|VX}(\cdot|\cdot, x_2)$. By Assumption 2, for each v the above admits a unique solution $f_{W|VX}(\cdot|\cdot, x_2)$. To see this, suppose that for a given v there are two solutions h_1 and h_2, both of which are bounded and integrable. Then we must have:

$$\int_{\mathcal{W}} f_{Z|W}(z|w)\big(h_1(w) - h_2(w)\big) dw = 0$$

But then we apply Assumption 2 with $\delta(w) = h_1(w) - h_2(w)$ and we see that $h_1(w) = h_2(w)$. Thus (A.1) identifies $f_{W|VX}(\cdot|\cdot, \cdot)$ up to reorderings of W. By similar reasoning we get that $f_{Y|WX}(\cdot|\cdot, x_2)$ is then identified from A.1 with x_1 repalced by x_2 up to reorderings of W.

Now, by elementary properties of probability densities we have:

$$f_{W|X}(w|x_2) = \int_{\mathcal{V}} f_{W|VX}(w|v, x_2) f_{V|X}(v|x_2) dv$$

The objects on the RHS above are all known (up to reorderings of W), and so $f_{W|X}$ is identified. Now note that by conclusion iv. of Proposition 1:

$$\begin{aligned} f_{Y(x_1)|X}(y|x_2) &= \int_{\mathcal{W}} f_{Y(x_1)|WX}(y|w, x_2) f_{W|X}(w|x_2) dw \\ &= \int_{\mathcal{W}} f_{Y(x_1)|WX}(y|w, x_1) f_{W|X}(w|x_2) dw \\ &= \int_{\mathcal{W}} f_{Y|WX}(y|w, x_1) f_{W|X}(w|x_2) dw \end{aligned}$$

The first equality follows by elementary properties of probability densities. The second equality follows from condition iv. Proposition 1 which implies $f_{Y(x_1)|WX}(y|w, x_2) = f_{Y(x_1)|WX}(y|w, x_1)$. The final equality uses the fact that $f_{Y(x_1)|WX}(y|w, x_1)$ equals to $f_{Y|WX}(y|w, x_1)$, this follows the 'consistency' property which states that $Y = Y(X)$. □

Proof of Theorem 2. Under conditions i. and ii. of Proposition 2, HS Assumption 3, Assumption 1, and Assumption 4, we can apply HS Theorem 1 with $C = X$. This yields identification of f_{XW}, $f_{V|W}$, and $f_{Z|W}$ up to reorderings of W. Now note that by elementary properties of probability densities:

$$f_{YZX}(y, z, x) = \int_{\mathcal{W}} f_{Y|WZX}(y|w, z, x) f_{Z|XW}(z|x, w) f_{XW}(x, w) dw$$

By conclusion ii of Proposition 2, $f_{Z|XW}(z|x,w) = f_{Z|W}(z|w)$. Condition iii. of Proposition 2 implies $f_{Y|XWZ}(y|x,w,z) = f_{Y|XW}(y|x,w)$. Substituting into the equation above we get:

$$f_{YZX}(y,z,x) = \int_{\mathcal{W}} f_{Y|WX}(y|w,x) f_{Z|X}(z|w) f_{XW}(x,w) dw$$

Other than $f_{Y|WX}$, all the objects in the above are already identified, at least up to reorderings of W. To show $f_{Y|WX}$ is the unique solution to the equation, note that by HS Assumption 3 there can be only one bounded integrable function h so that for all $z \in \mathcal{Z}$:

$$f_{YZX}(y,z,x) = \int_{\mathcal{W}} h(w) f_{Z|W}(z|w) dw$$

If there were two such functions, h_1 and h_2 then we would have:

$$\int_{\mathcal{W}} \big(h_1(w) - h_2(w)\big) f_{Z|W}(z|w) dw = 0, \, \forall z \in \mathcal{Z}$$

and so by HS Assumption 3, $h_1(W) - h_2(W) = 0$ almost surely. Therefore, the only solution is:

$$h(w) = f_{Y|WX}(y|w,x) f_{XW}(x,w)$$

Since f_{XW} is non-zero by Assumption 1, there is a unique $f_{Y|WX}$ that satisfies the above. Thus $f_{Y|WX}$ is identified up to reorderings of W. $f_{W|X}$ is identified because $f_{W|X}(w|x) = f_{XW}(x,w)/f_X(x)$. Under conclusion iv. of Proposition 2 we can now apply the final steps in the proof of Theorem 1 to identify the conditional and marginal distributions of potential outcomes. □

Proof of Theorem 3. Under conditions i. and ii. of Proposition 3 and Assumption 1, 5, and 6, we can apply HS Theorem 1 conditional on $Y = y$ with $C = X$. This yields identification of $f_{X|WY}(\cdot|\cdot,y)$, $f_{Z|WY}(\cdot|\cdot,y)$, and $f_{W|VY}(\cdot|\cdot,y)$ (up to reorderings of W) from the equation below:

$$f_{XZ|VY}(x,z|v,y) = \int_{\mathcal{W}} f_{X|WY}(x|w,y) f_{Z|W}(z|w) f_{W|VY}(w|v,y) dw \qquad \text{(A.2)}$$

In the above we have imposed condition iii. of Proposition 3 which implies that $f_{Z|WY}(z|w,y) = f_{Z|W}(z|w)$. Taking (A.2) with y' in place of y and integrating over $x \in \mathcal{X}$ we get:

$$f_{Z|VY}(z|v,y') = \int_{\mathcal{W}} f_{Z|W}(z|w) f_{W|VY}(w|v,y') dw$$

Other than $f_{W|VY}(\cdot|\cdot,y')$, all the objects in the equation above are identified (at least up to reorderings of W). By Assumption 5, for each z the equation above has a unique solution $f_{W|VY}(\cdot|\cdot,y')$ (see the reasoning in the proof of Theorem 1). Thus we have identified $f_{W|VY}(\cdot|\cdot,\cdot)$ up to reorderings of W from (A.2). By

similar reasoning $f_{X|WY}(x|w,y')$ is identified from (A.2) up to reorderings of W. Now, by elementary properties of densities:

$$f_{W|X}(w|x) = \int_{\mathcal{Y}} \int_{\mathcal{V}} f_{X|WY}(x|w,y) f_{W|VY}(w|v,y) \frac{f_{VY}(v,y)}{f_X(x)} dv dy$$

So $f_{W|X}$ is identified up to reorderings of W. Applying conclusions i. and ii. of Proposition 3 and elementary properties of densities, we get:

$$\begin{aligned} f_{YZ|X}(y,z|x) &= \int_{\mathcal{W}} f_{Z|XYW}(z|x,y,w) f_{YW|X}(y,w|x) dw \\ &= \int_{\mathcal{W}} f_{Z|YW}(z|w) f_{Y|WX}(y|w,x) f_{W|X}(w|x) dw \end{aligned}$$

$f_{V|YW}(v|y,w)$ and $f_{W|X}(w|x)$ are known up to reorderings of W. By Assumptions 1 and 5, for each y and x there is a unique solution $f_{Y|WX}(y|\cdot,x)$ to the above (see the reasoning in the proof of Theorem 2), and so this equation identifies $f_{Y|WX}$ up to reorderings of W.

Finally, we apply conclusion iv. of Proposition 3 and follow the same steps as in the proof of Theorem 1 to identify the conditional and marginal distributions of potential outcomes. □

Proof of Theorem 4. Under conclusions i. and ii. of proposition 4 and Assumptions 2, 7, and 8 we can apply HS Theorem 1 within the stratum x_1 of X. This yields identification of $f_{C|WX}(\cdot|\cdot,x_1)$, $f_{W|VX}(\cdot|\cdot,x_1)$, and $f_{Z|WX}(\cdot|\cdot,x_1)$ up to reorderings of W rom the equation below:

$$f_{CZ|VX}(c,z|v,x_1) = \int_{\mathcal{W}} f_{C|WX}(c|w,x_1) f_{W|VX}(w|v,x_1) f_{Z|W}(z|w) dw$$

Note we have imposed $f_{Z|WX}(z|w,x_1) = f_{Z|W}(z|w)$ which follows from conclusion iii. of Proposition 4. Taking the equation above with x_1 replaced by x_2 and integrating over c we get:

$$f_{Z|VX}(z|v,x_2) = \int_{\mathcal{W}} f_{W|VX}(w|v,x_2) f_{Z|W}(z|w) dw$$

The only object in the above that is not already identified up to orderings of W is $f_{W|VX}(\cdot|\cdot,x_2)$. Assumption 2 implies there is a unique solution $f_{W|VX}(\cdot|\cdot,x_2)$ to the above (see the proof of Theorems 1) and so $f_{W|VX}(\cdot|\cdot,\cdot)$ is identified up to reorderings of W. Similar reasoning then identifies $f_{C|WX}$ up to reorderings of W.

Now, by elementary properties of probability densities we have:

$$f_{YZ|VX}(y,z|v,x) = \int_{\mathcal{W}} f_{Y|WVZX}(y|w,v,z,x) f_{Z|WVX}(z|w,v,x) f_{W|VX}(w|v,x) dw$$

Proposition 4.iv implies $f_{Y|WVZX}(y|w,v,z,x) = f_{Y|WVX}(y|w,v,x)$. In addition, conclusions ii. and iii. of Proposition 4 imply $(V,X) \perp\!\!\!\perp Z|W$ and so

$f_{Z|WVX}(z|w,v,x) = f_{Z|W}(z|w)$. Substituting and again applying elementary properties of probability densities we get:

$$f_{YZ|VX}(y,z|v,x) = \int_{\mathcal{W}} f_{Y|WVX}(y|w,v,x) f_{Z|W}(z|w) f_{W|VX}(w|v,x) dw$$

All objects in the equation above are identified up to reorderings of W, other than $f_{Y|WVX}$. By the same reasoning as in the proof of Theorem 1 Assumption 2 implies there is only one bounded integrable function h so that for all $z \in \mathcal{Z}$:

$$f_{YVZ|X}(y,v,z|x) = \int_{\mathcal{W}} h(w) f_{Z|WX}(z|w,x) dw$$

The only solution to the above is:

$$h(w) = f_{Y|WVX}(y|w,v,x) f_{W|VX}(w|v,x)$$

$f_{W|VX}(w|v,x)$ is non-zero by Assumption 7, so there is a unique $f_{Y|WVX}$ that satisfies the above. Thus we have identified $f_{Y|WVX}$ up to reorderings of W. Finally, using conclusion v. of Proposition 4 and elementary properties of densities:

$$\begin{aligned}
f_{Y(x_1)|X}(y|x_2) &= \int_{\mathcal{W}} \int_{\mathcal{V}} f_{Y(x_1)|WVX}(y|w,v,x_2) f_{W|VX}(w|v,x_2) f_{V|X}(v|x_2) dv dw \\
&= \int_{\mathcal{W}} \int_{\mathcal{V}} f_{Y(x_1)|WV}(y|w,v,x_1) f_{W|VX}(w|v,x_2) f_{V|X}(v|x_2) dv dw \\
&= \int_{\mathcal{W}} \int_{\mathcal{V}} f_{Y|WVX}(y|w,v,x_1) f_{W|VX}(w|v,x_2) f_{V|X}(v|x_2) dv dw
\end{aligned}$$

Where the second equality uses $f_{Y(x_1)|WVX}(y|w,v,x_2) = f_{Y(x_1)|WVX}(y|w,v,x_1)$ by Proposition 4.v, and for the final equality we use $f_{Y(x_1)|WV}(y|w,v,x_1) = f_{Y|WVX}(y|w,v,x_1)$ which follows by consistency. $f_{Y(x_1)}$ is then identified from $f_{Y(x_1)|X}$ as in Theorem 1. □

Proof of Lemma 1. By consistency $f_{Y|XW}(y|x,w) = f_{Y(x,w)|XW}(y|x,w)$. By conclusion i. of Proposition 5, $f_{Y(x,w)|XW}(y|x,w) = f_{Y(x,w)}(y)$. Combining gives $f_{Y|XW}(y|x,w) = f_{Y(x,w)}(y)$. Also by consistency $f_{Y|XVW}(y|x,v,w) = f_{Y(x,w)|XVW}(y|x,v,w)$, and by conclusion ii. of Proposition 5, $f_{Y(x,w)|V}(y|v) = f_{Y(x,w)|XVW}(y|x,v,w)$. Combining gives $f_{Y(x,w)|V}(y|v) = f_{Y|XVW}(y|x,v,w)$. Multiplying both sides by $f_V(v)$ and integrating over $v \in \mathcal{V}$ we get:

$$f_{Y(x,w)}(y) = \int_{\mathcal{V}} f_{Y|XVW}(y|x,v,w) f_V(v) dv$$

□

Proof of Theorem 5. Combining Lemma 1 and any of Theorems 1, 2, 3, or 4 we achieve identification of $f_{Y(x,w)}$, $f_{Z|W}$, and either $f_{W|X}$ or $f_{W|VX}$ up to reorderings of W. Note that the latter implies identification of f_W up to reorderings

of W. More precisely, $\tilde{f}_{Y(x,w)}$, $\tilde{f}_{Z|W}$, and $\tilde{f}_W$ are identified, where $\tilde{f}_{Y(x,w)}(y) = f_{Y(x,\varphi^{-1}(w))}(y)$, $\tilde{f}_{Z|\varphi(W)}(z|w) = f_{Z|W}(z|w)$, and $\tilde{f}_W(w) = f_{\varphi(W)}(w)$ for an unknown injective function φ. Define a function α by:

$$\alpha(w) = M[\tilde{f}_{Z|W}(\cdot|w)]$$

Under HS Assumption 5, $M[f_{Z|W}(\cdot|w)] = w$, this implies that:

$$\alpha(w) = \varphi^{-1}(w)$$

Since M is known and $\tilde{f}_{Z|W}$ is identified, the equation above identifies the function φ^{-1}. $f_{Y(x,w)}$ is then identified from:

$$f_{Y(x,w)}(y) = \tilde{f}_{Y(x,\alpha^{-1}(w))}(y)$$

Under Assumption 9, which is weaker than HS Assumption 5, there is a component-wise strictly increasing function ϕ so that $M[f_{Z|W}(\cdot|w)] = \phi(w)$, and so:

$$\alpha(w) = \phi(\varphi^{-1}(w))$$

Let $\tilde{f}_{\alpha(W)}$ be the density of $\alpha(\tilde{W})$ where $\tilde{W}$ is a random variable with density $\tilde{f}_W$. The equation above implies that:

$$\tilde{f}_{\alpha(W)}(w) = f_{\phi(W)}(w)$$

Note that $\tilde{f}_{\alpha(W)}$ is identified because α and $\tilde{f}_W$ are identified. Since ϕ is strictly increasing $Q_{\phi(W)}(\tau) = \phi\big(Q_W(\tau)\big)$. Let $Q_{\alpha(\tilde{W})}(\tau)$ be the quantiles of $\alpha(\tilde{W})$, then we have $Q_{\alpha(\tilde{W})}(\tau) = \phi\big(Q_W(\tau)\big)$ and $\alpha^{-1}\big(Q_{\alpha(\tilde{W})}(\tau)\big) = \varphi\big(Q_W(\tau)\big)$. Using $\tilde{f}_{Y(x,w)}(y) = f_{Y(x,\varphi^{-1}(w))}(y)$ we then get:

$$\tilde{f}_{Y\left(x,\alpha^{-1}(Q_{\alpha(\tilde{W})}(\tau))\right)}(y) = f_{Y\left(x,Q_W(\tau)\right)}(y)$$

□

Lemma 2. *Suppose X is binary, Assumption 10 holds, and for $x = 0, 1$, $f_{Y(x)|W}$ is identified up to reorderings of W which can differ for each x. Define $\bar{s}$ and $\underline{s}$ as in the statement of Theorem 6. Then for almost all $w \in \mathcal{W}$:*

$$E[Y(1) - Y(0)|W = w] \in [\underline{s}, \bar{s}]$$

Proof. By supposition, for each $x \in \mathcal{X}$, $f_{Y(x)|W}$ is identified up to reorderings of W which can differ for each x. More formally, we identify a function $\tilde{f}_{Y(x)|W}$ so that there is an unknown function $\varphi(w, x)$ so that $\varphi(\cdot, x)$ is injective with inverse $\varphi^{-1}(\cdot, x)$ for each x, and $\tilde{f}_{Y(x)|W}(y|w) = f_{Y(x)|\varphi(W,X)}(y|w)$. Note then that:

$$\int_{\mathcal{Y}} y\tilde{f}_{Y(x)|W}(y|w)dy = E[Y(x)|W = \varphi^{-1}(w, x)]$$

The above implies that:

$$\operatorname*{ess\,sup}_{w\in\mathcal{W}} \int_{\mathcal{Y}} y\tilde{f}_{Y(x)|W}(y|w)dy = \operatorname*{ess\,sup}_{w\in\mathcal{W}} E[Y(x)|W=w]$$

And similarly for the essential infima. Substituting into the definitions of $\bar{s}$ and $\underline{s}$ we get:

$$\bar{s} = \operatorname*{ess\,sup}_{w\in\mathcal{W}} E[Y(1)|W=w] - \operatorname*{ess\,sup}_{w\in\mathcal{W}} E[Y(0)|W=w]$$
$$\underline{s} = \operatorname*{ess\,inf}_{w\in\mathcal{W}} E[Y(1)|W=w] - \operatorname*{ess\,inf}_{w\in\mathcal{W}} E[Y(0)|W=w]$$

Next we will use the above to show that, under monotonicity, $\bar{s}$ is the supremum of the CATE and $\underline{s}$ is the infimum. Let $\{w_n\}_{n=1}^{\infty}$ be a sequence in $\mathcal{W}$ so that:

$$E[Y(0)|W=w_n] \to \operatorname*{ess\,sup}_{w\in\mathcal{W}} E[Y(0)|W=w]$$

The above implies that:

$$P(E[Y(0)|W] \leq E[Y(0)|W=w_n]) \to 1$$

By monotonicity, if $E[Y(0)|W=w] \leq E[Y(0)|W=w_n]$ then:

$$E[Y(1)-Y(0)|W=w] \leq E[Y(1)-Y(0)|W=w_n]$$

And so:

$$P(E[Y(1)-Y(0)|W] \leq E[Y(1)-Y(0)|W=w_n]) \to 1$$

The above then implies:

$$E[Y(1)-Y(0)|W=w_n] \to \operatorname*{ess\,sup}_{w\in\mathcal{W}} E[Y(1)-Y(0)|W=w]$$

Also by monotonicity, $E[Y(0)|W=w] \leq E[Y(0)|W=w_n]$ implies that $E[Y(1)|W=w] \leq E[Y(1)|W=w_n]$, so by similar reasoning we get:

$$E[Y(1)|W=w_n] \to \operatorname*{ess\,sup}_{w\in\mathcal{W}} E[Y(1)|W=w]$$

Now, using linearity of the expectation we have:

$$\begin{aligned}\lim_{n\to\infty} \big(E[Y(1)|W=w_n] - E[Y(0)|W=w_n]\big) &= \lim_{n\to\infty} E[Y(1)-Y(0)|W=w_n] \\ &= \operatorname*{ess\,sup}_{w\in\mathcal{W}} E[Y(1)-Y(0)|W=w]\end{aligned}$$

By Assumption 10, $\operatorname*{ess\,sup}_{w\in\mathcal{W}} E[Y(0)|W=w] \leq c < \infty$ and so we have:

$$\begin{aligned}&\lim_{n\to\infty} \big(E[Y(1)|W=w_n] - E[Y(0)|W=w_n]\big) \\ =&\lim_{n\to\infty} E[Y(1)|W=w_n] - \lim_{n\to\infty} E[Y(0)|W=w_n] \\ =&\bar{s}\end{aligned}$$

Combining we get:

$$\bar{s} = \operatorname*{ess\,sup}_{w \in \mathcal{W}} E[Y(1) - Y(0)|W = w]$$

Following similar steps we get:

$$\underline{s} = \operatorname*{ess\,inf}_{w \in \mathcal{W}} E[Y(1) - Y(0)|W = w]$$

It now follows immediately from the definition of the essential supremum and infimum that for almost all $w \in \mathcal{W}$:

$$E[Y(1) - Y(0)|W = w] \in [\underline{s}, \bar{s}]$$

□

Proof of Theorem 6. Under conditions i. and ii. of Proposition 6 and Assumption 1, 2, and 3, for each $x \in \mathcal{X}$ we can apply HS Theorem 1 conditional on $X = x$ with $C = Y$. This yields identification of $f_{Y|WX}(\cdot|\cdot, x)$, $f_{Z|WX}(\cdot|\cdot, x)$, and $f_{W|VX}(\cdot|\cdot, x)$ up to reorderings of W which may vary with x. Now note that by condition iii. of Proposition 6, $f_{Y|WX}(y|w, x) = f_{Y(x)|W}(y|w)$. So $f_{Y(x)|W}$ is identified up to reorderings of W which may depend on x. We then apply Lemma 2 to get that for almost all $w \in \mathcal{W}$:

$$E[Y(1) - Y(0)|W = w] \in [\underline{s}, \bar{s}]$$

For the final result in the theorem note that under conclusion iii. of Proposition 6 we have $E[Y(x)|W = w] = E[Y(x)|W = w, X = x]$ for all $x \in \mathcal{X}$, and so:

$$E[Y(1) - Y(0)|W = w] = E[Y(1) - Y(0)|W = w, X = x]$$

Applying the law of iterated expectations:

$$E[Y(1) - Y(0)|X = x] = E\big[E[Y(1) - Y(0)|W]|X = x\big]$$

We have established that with probability 1, $E[Y(1) - Y(0)|W, X] \in [\underline{s}, \bar{s}]$ and thus the same holds for the conditional mean of this random variable.

□

Proof of Theorem 7. Under conclusions i. ii., and iii. of Proposition 7 and Assumptions 2, 7, and 8 we can apply HS Theorem 1 within the stratum x of X. This yields identification of $f_{C|WX}(\cdot|\cdot, x)$, $f_{W|VX}(\cdot|\cdot, x)$, and $f_{Z|WX}(\cdot|\cdot, x)$ up to reorderings of W which may depend on x. Now note that by elementary properties of probability densities:

$$\begin{aligned} & f_{YZ|VX}(y, z|v, x) \\ = & \int_{\mathcal{W}} f_{Y|WVZX}(y|w, v, z, x) f_{Z|WVX}(z|w, v, x) f_{W|VX}(w|v, x) dw \end{aligned}$$

Proposition 7.iii implies $f_{Y|WVZX}(y|w,v,z,x) = f_{Y|WVX}(y|w,v,x)$. In addition, conclusion ii. of Proposition 7 implies $f_{Z|WVX}(z|w,v,x) = f_{Z|WX}(z|w,x)$. Substituting we get:

$$f_{YZ|VX}(y,z|v,x) = \int_{\mathcal{W}} f_{Y|WVX}(y|w,v,x) f_{Z|WX}(z|w,x) f_{W|VX}(w|v,x) dw$$

By Assumption 2 the expression above identifies $f_{Y|WVX}(\cdot|\cdot,\cdot,x)$ up to reorderings of W (see the steps in the proof of Theorem 4). Now note that under conclusion iv. of Proposition 7 we get:

$$f_{Y|WVX}(y|w,v,x) = f_{Y(x)|WV}(y|w,v)$$

So $f_{Y(x)|WV}(y|w,v)$ is identified up to reorderings of W which may depend on x.

We now apply Lemma 2 within each stratum v of V, using Assumption 11 in place of Assumption 10. We get that for almost all $w \in \mathcal{W}$ and $v \in \mathcal{V}$:

$$E[Y(1) - Y(0)|W = w, V = v] \in [\underline{s}(v), \bar{s}(v)]$$

Finally, note that under conclusion iv. of Proposition 7 we have that for all $x \in \mathcal{X}$ and almost all $w \in \mathcal{W}$ and $v \in \mathcal{V}$

$$E[Y(x)|W = w, V = v] = E[Y(x)|W = w, V = v, X = x]$$

Using the above and applying the law of iterated expectations we get:

$$E[Y(1) - Y(0)|X = x] = E\big[E[Y(1) - Y(0)|W, V]|X = x\big]$$

Since $E[Y(1)-Y(0)|W,V] \leq \bar{s}(V)$ almost surely we have $E\big[E[Y(1)-Y(0)|W,V]|X = x\big] \leq E[\bar{s}(V)|X = x]$ and similarly for the lower bound. □