I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11866

A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands

Tymon Słoczyński

OCTOBER 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11866

A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands

Tymon Słoczyński
Brandeis University and IZA

OCTOBER 2018

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands*

It is standard practice in applied work to study the effect of a binary variable ("treatment") on an outcome of interest using linear models with additive effects. In this paper I study the interpretation of the ordinary and two-stage least squares estimands in such models when treatment effects are in fact heterogeneous. I show that in both cases the coefficient on treatment is identical to a convex combination of two other parameters (different for OLS and 2SLS), which can be interpreted as the average treatment effects on the treated and controls under additional assumptions. Importantly, the OLS and 2SLS weights on these parameters are inversely related to the proportion of each group. The more units get treatment, the less weight is placed on the effect on the treated. What follows, the reliance on these implicit weights can have serious consequences for applied work. I illustrate some of these issues in four empirical applications from different fields of economics. I also develop a weighted least squares correction and simple diagnostic tools that applied researchers can use to avoid potential biases. In an important special case, my diagnostics only require the knowledge of the proportion of treated units.

JEL Classification: C21, C24, C26, C31

Keywords: heterogeneity, ordinary least squares, propensity score, two-stage least squares, treatment effects

Corresponding author:
Tymon Słoczyński
Department of Economics
Brandeis University
415 South Street, MS 021
Waltham, MA 02453
USA

E-mail: tslocz@brandeis.edu

* I am particularly grateful for helpful comments and discussions with Alberto Abadie, Max Kasy, Pedro Sant'Anna, and Jeff Wooldridge. I also thank Arun Advani, Isaiah Andrews, Josh Angrist, Orley Ashenfelter, Richard Blundell, Stéphane Bonhomme, Carol Caetano, Marco Caliendo, Matias Cattaneo, Gary Chamberlain, Todd Elder, Alfonso Flores-Lagunes, Andreas Hagemann, James Heckman, Kei Hirano, Macartan Humphreys, Guido Imbens, Krzysztof Karbownik, Toru Kitagawa, Pat Kline, Paweł Królikowski, Nicholas Longford, James MacKinnon, Łukasz Marć, Doug Miller, Michał Myck, Mateusz Myśliwski, Gary Solon, Jann Spiess, Michela Tincani, Alex Torgovitsky, Joanna Tyrowicz, Takuya Ura, Rudolf Winter-Ebmer, seminar participants at Brandeis, Harvard–MIT, Holy Cross, IHS Vienna, Lehigh, Michigan State, Potsdam, SDU Odense, SGH, Temple, UCL, Upjohn, WZB Berlin, and participants at many conferences for useful feedback. I acknowledge financial support from the National Science Centre (grant DEC-2012/05/N/HS4/00395), the Foundation for Polish Science (a "Start" scholarship), the "Weź stypendium – dla rozwoju" scholarship program, and the Theodore and Jane Norman Fund. This paper builds on ideas from and supersedes my paper "New Evidence on Linear Regression and Treatment Effect Heterogeneity" (IZA DP No. 9491).

1 Introduction

Many applied researchers study the effect of a binary variable ("treatment") on the expected value of some outcome of interest, holding fixed a vector of other covariates. As noted by Imbens (2015), despite the availability of a large number of semi- and nonparametric estimators for average treatment effects, applied researchers typically continue to use conventional regression methods. In particular, it is standard practice in applied work to use ordinary least squares (OLS) to estimate

$$y = \alpha + \tau d + X\beta + u, \tag{1}$$

where y denotes the outcome, d denotes the binary variable of interest, and X denotes the row vector of other covariates (control variables); τ is then usually interpreted as the average treatment effect (ATE). This simple estimation strategy is used in a large number of applied papers in leading economics journals, as well as in other disciplines.[1] Similarly, when applied researchers have access to a vector of instrumental variables (IV) for d (denoted by Z), it is standard practice to estimate the model in (1) using two-stage least squares (2SLS). Such an estimation strategy is also widespread in applied work.[2]

The great appeal of the model in (1) comes from its simplicity (see, *e.g.*, Angrist and Pischke, 2009). At the same time, however, a large body of evidence demonstrates the empirical importance of heterogeneity in effects (see, *e.g.*, Heckman, 2001; Bitler, Gelbach, and Hoynes, 2006, 2008) which is explicitly ruled out by this same model. In this paper, therefore, I study the interpretation of the OLS and 2SLS estimands in (1) when treatment effects are in fact heterogeneous. I derive a new theoretical result which demonstrates that both estimands are identical to the outcome of the following three-step procedure: in the first step, take the linear projection of d on X (OLS) or the linear projection of d on X and the first-stage errors (2SLS), which gives the "propensity score" from the

[1] See, *e.g.*, Black, Smith, Berger, and Noel (2003), Almond, Chay, and Lee (2005), Campbell, Giglio, and Pathak (2011), Voigtländer and Voth (2012), Alesina, Giuliano, and Nunn (2013), Berger, Easterly, Nunn, and Satyanath (2013), Frakes (2013), Martinez-Bravo (2014), Aizer, Eli, Ferrie, and Lleras-Muney (2016), Atkin (2016), Bobonis, Cámara Fuertes, and Schwabe (2016), Das, Holla, Mohpal, and Muralidharan (2016), Michalopoulos and Papaioannou (2016), and Belenzon, Chatterji, and Daley (2017).

[2] Following Imbens and Angrist (1994), the 2SLS estimand is usually interpreted as the local average treatment effect (LATE). In principle, however, such an interpretation is not accurate, because the result in Imbens and Angrist (1994) applies only to models without covariates (X). For results on models with covariates, see Angrist and Imbens (1995) and Abadie (2003). For recent applications of this strategy, see Dinkelman (2011), Dittmar (2011), Parey and Waldinger (2011), Jacob and Ludwig (2012), Maestas, Mullen, and Strand (2013), Deming, Hastings, Kane, and Staiger (2014), Moser, Voena, and Waldinger (2014), Black, Sanders, Taylor, and Taylor (2015), Dobbie and Song (2015), Clark and Del Bono (2016), Bettinger, Fox, Loeb, and Taylor (2017), Lundborg, Plug, and Rasmussen (2017), and many others.

linear probability model; in the second step, project y on d, the propensity score, and their interaction—and calculate average partial effects from this model for both groups of interest ("treated" and "controls"); in the third step, calculate a weighted average of these two effects—with weights being *inversely related* to the unconditional probability that a unit belongs to a given group.[3] In consequence, when the proportion of one group *increases*, the weight on the effect on this group *decreases*. I also establish conditions under which either of these estimation strategies recovers

$$\tau = \mathrm{P}\left(d=1\right) \cdot \tau_{ATC} + \mathrm{P}\left(d=0\right) \cdot \tau_{ATT} \tag{2}$$

instead of

$$\tau_{ATE} = \mathrm{P}\left(d=1\right) \cdot \tau_{ATT} + \mathrm{P}\left(d=0\right) \cdot \tau_{ATC}, \tag{3}$$

where τ_{ATE} denotes the average treatment effect (ATE), τ_{ATT} denotes the average treatment effect on the treated (ATT), and τ_{ATC} denotes the average treatment effect on the controls (ATC); also, $\mathrm{P}\left(d=1\right)$ and $\mathrm{P}\left(d=0\right)$ denote population proportions of treated and control units, respectively. As a consequence of the disparity between (2) and (3), in many empirical applications the ordinary and two-stage least squares estimands might not be close to any of the average treatment effects of interest.

On the one hand, the implications of this result are necessarily pessimistic. It might seem sensible to resort to alternative methods which explicitly allow for treatment effect heterogeneity. On the other hand, it is possible to approach this result from a more pragmatic perspective. In particular, in this paper I also develop diagnostic tools which can detect deviations of the OLS/2SLS weights from the pattern which would be consistent with the chosen target parameter, ATE or ATT. These diagnostics are easy to implement and interpret. When one of my diagnostics is close to zero, OLS or 2SLS estimation of the model in (1) is likely to provide a reasonable estimate of the corresponding target parameter; if this diagnostic gets closer to one in absolute value, other methods are probably more appropriate. Interestingly, in an important special case, these diagnostics only depend on the proportion of treated units. When this proportion is close to zero (one), the OLS/2SLS estimates will be similar to those of the effect on the treated (controls); when this same proportion is close to 50 percent, the OLS/2SLS estimates will be similar to the implicit estimates of the average treatment effect.

I also extend my baseline results in several other directions. In particular, I develop a simple weighted least squares correction which is capable of undoing the OLS/2SLS

[3] The second and third steps are identical for OLS and 2SLS, with the only difference in the first-step calculation of the propensity scores.

weights. This procedure automatically recovers the implicit estimate of the average treatment effect. Also, I discuss the implications of my main result for regression adjustments to experimental data, fixed effects, and difference-in-differences estimation. Finally, I explicitly address the common interpretation of the IV estimand as the local average treatment effect, and briefly discuss conditions under which this interpretation might be accurate even in the presence of other covariates.

This paper contributes to a growing field of research in econometrics which studies weighted average representations of various estimands and estimators.[4] My theoretical results are most closely related to one line of such research, concerning both OLS and 2SLS, which investigates the interpretation of the coefficient on treatment when the model for covariates is saturated (Angrist and Imbens, 1995; Angrist, 1998; Humphreys, 2009), *i.e.* the estimating equation includes a binary variable for each combination of covariate values ("stratum"). In the case of 2SLS, the first stage also needs to be fully saturated (Angrist and Imbens, 1995). In this restricted setting, Angrist and Imbens (1995) demonstrate that the weights underlying two-stage least squares are proportional to the variance of treatment in each stratum. Angrist (1998) provides an analogous result for ordinary least squares.[5] Humphreys (2009) extends this latter result and shows that the OLS estimand is bounded by τ_{ATT} and τ_{ATC} whenever treatment assignment probabilities are monotonic in stratum-specific effects. What is particularly restrictive, however, is the implicit requirement of each of these papers that all of the covariates and instruments are discrete, since otherwise saturation would not be possible.

What sets this paper apart from these previous contributions is its much wider applicability. Most importantly, I relax the saturated model restriction—which is rarely used

[4] See, *e.g.*, Yitzhaki (1996), Deaton (1997), Angrist (1998), Angrist and Krueger (1999), Humphreys (2009), Løken, Mogstad, and Wiswall (2012), Solon, Haider, and Wooldridge (2015), and Kato and Sasaki (2017) for studies of OLS; Imbens and Angrist (1994), Angrist and Imbens (1995), Angrist, Imbens, and Rubin (1996), Angrist, Graddy, and Imbens (2000), Abadie (2003), Løken *et al.* (2012), Kolesár (2013), Andrews (2017), and Evdokimov and Kolesár (2018) for studies of IV methods; Wooldridge (2005), Løken *et al.* (2012), Chernozhukov, Fernández-Val, Hahn, and Newey (2013), Imai and Kim (2017), and Gibbons, Suárez Serrato, and Urbancic (2018) for studies of fixed effects; Borusyak and Jaravel (2017), Abraham and Sun (2018), Athey and Imbens (2018), de Chaisemartin and D'Haultfœuille (2018), Goodman-Bacon (2018), Hull (2018), and Strezhnev (2018) for studies of difference-in-differences and related methods; and Kato and Sasaki (2017) for a study of quantile regression. This literature is also related to Heckman and Vytlacil (2005), Heckman, Urzua, and Vytlacil (2006), and Heckman and Vytlacil (2007) who provide an interpretation of various estimands, *conditional on* X, as weighted averages of marginal treatment effects.

[5] A similar result for nonsaturated models is derived by Rhodes (2010) and Aronow and Samii (2016). In both of these papers the OLS estimand is interpreted as a weighted average of individual-level treatment effects. In this paper I provide an alternative representation, in which this estimand is interpreted as a weighted average of group-specific average treatment effects (τ_{ATT} and τ_{ATC}). This representation makes it much easier to distinguish between applications in which OLS weighting might or might not be problematic. Indeed, in this paper I develop simple diagnostic tools based on my main result.

in applied work—and I do not require the covariates or instruments to be discrete. Also, instead of focusing on stratum-specific treatment effects, I provide a general weighted average representation of both the ordinary and two-stage least squares estimands—in terms of group-specific average treatment effects (τ_{ATT} and τ_{ATC}). This formulation is very attractive—and easier to interpret—because each OLS or 2SLS estimate can now be expressed as a weighted average of two estimates of τ_{ATT} and τ_{ATC}. Moreover, the weights are also easily estimated—and they are always nonnegative and sum to one.

To illustrate the importance of this result, I also replicate several prominent applied studies. My starting point is to outline various scenarios that applied researchers are likely to face in the context of OLS/2SLS estimation and treatment effect heterogeneity. These scenarios are defined by our parameter of interest (ATE or ATT), the proportion of treated units, and the amount of heterogeneity. In one scenario, we are only interested in one of the parameters and my diagnostics suggest that this parameter will be recovered even if treatment effects are highly heterogeneous. In another scenario, my diagnostics suggest caution in interpreting the OLS/2SLS estimates but this concern is outweighed by the small amount of heterogeneity. Next, we are interested in both parameters and my diagnostics suggest that one of them—but of course not both—is likely to be recovered by the OLS/2SLS estimates regardless of whether treatment effects are heterogeneous or not. Finally, my diagnostics might also suggest substantial bias in estimating our target parameter or parameters in the presence of substantial heterogeneity.

In this paper I present an empirical illustration for each of these scenarios. In particular, I replicate the analysis of the effects of the National Supported Work (NSW) program in Angrist and Pischke (2009); the study of the association of medieval pogroms with modern anti-Semitism in Voigtländer and Voth (2012); the analysis of the impact of Catholic schooling on math test scores in Wooldridge (2015); and the study of the long-run effects of cash transfers in Aizer *et al.* (2016). These empirical applications confirm the importance of the theoretical results presented in this paper. In particular, they confirm the usefulness of my diagnostic methods in distinguishing between cases where OLS or 2SLS estimation of the model in (1) will and will not be problematic.

2 Theory

This section contains the main methodological results of this paper. After presenting a simple numerical example and introducing the basic concepts, I discuss my main contribution, namely a weighted average representation of the ordinary least squares estimand. I provide the intuition behind this result, illustrate it graphically, and derive a number of useful corollaries. Of particular interest is the development of simple diagnostics for undesirable weighting of heterogeneous treatment effects in least squares estimation. These diagnostics are easy to implement and interpret; in an important special case, their calculation only requires the knowledge of the proportion of treated units. Finally, I develop a weighted least squares procedure which can be used to undo the OLS weighting and I also extend my main result to two-stage least squares. Other theoretical extensions are relegated toward the end of the paper.

Notation

Throughout this paper, scalar random variables are denoted by lowercase letters and vectors of random variables are denoted by uppercase letters. In particular, let y denote the outcome, let d denote the binary variable of interest ("treatment"), and let X denote the row vector of other covariates, $(x_1, \ldots, x_K)$. There are two potential outcomes, $y(1)$ and $y(0)$, but we observe only one of them for each unit, $y = y(d) = y(1) \cdot d + y(0) \cdot (1 - d)$. The usual parameters of interest are $\tau_{ATE} = \mathrm{E}\left[y(1) - y(0)\right]$, $\tau_{ATT} = \mathrm{E}\left[y(1) - y(0) \mid d = 1\right]$, and $\tau_{ATC} = \mathrm{E}\left[y(1) - y(0) \mid d = 0\right]$.

Motivating Example

To illustrate the potential problem with OLS weighting, consider the following example with a single binary covariate, x. Treatment is unconfounded conditional on x. The researcher regresses y on d and x, ignoring the interaction term $d \cdot x$. Let $\mathrm{P}(d = 1) = 4\%$, $\mathrm{P}(x = 1 \mid d = 1) = 50\%$, and $\mathrm{P}(x = 1 \mid d = 0) = 18.75\%$. While the proportion of treated units is rather small at 4%, this is not unusual in the program evaluation literature; it will also help to make the illustration more effective. The information so far is sufficient to obtain the OLS weights on $\tau_1 = \mathrm{E}\left[y(1) - y(0) \mid x = 1\right]$ and $\tau_0 = \mathrm{E}\left[y(1) - y(0) \mid x = 0\right]$, as derived by Angrist (1998).[6] Indeed, simple algebra shows that the weight on τ_1 is 48% and the weight on τ_0 is 52%. Also, $\mathrm{P}(x = 1) = 20\%$. In other words, the weight on τ_1, equal to 48%, is much larger than we would otherwise expect (*i.e.*, 20%), but the result in Angrist (1998) justifies this by the fact that $\mathrm{V}(d \mid x = 1) > \mathrm{V}(d \mid x = 0)$. While it is clear that the OLS estimand will be different from τ_{ATE}, it is not as straightforward to realize that this difference will be functionally related to $\mathrm{P}(d = 1)$.

[6] The weight on τ_1 is equal to $\frac{\mathrm{P}(x=1) \cdot \mathrm{V}(d|x=1)}{\mathrm{P}(x=0) \cdot \mathrm{V}(d|x=0) + \mathrm{P}(x=1) \cdot \mathrm{V}(d|x=1)}$ and the weight on τ_0 is equal to $\frac{\mathrm{P}(x=0) \cdot \mathrm{V}(d|x=0)}{\mathrm{P}(x=0) \cdot \mathrm{V}(d|x=0) + \mathrm{P}(x=1) \cdot \mathrm{V}(d|x=1)}$. See also Angrist (1998) and Angrist and Pischke (2009) for more details.

Table 1: Numerical Example

	Weight on τ_1	Weight on τ_0	Weight on τ_{ATT}	Weight on τ_{ATC}	Parameter value
OLS	48%	52%	93.6%	6.4%	0.92
τ_{ATE}	20%	80%	4%	96%	–0.2
τ_{ATT}	50%	50%	100%	0%	1
τ_{ATC}	18.75%	81.25%	0%	100%	–0.25

Notes: All values reported in this table correspond to the numerical example discussed in the main text. In this example, there is a single binary covariate, x. "OLS" is the coefficient on d in the regression of y on d and x. The following information is sufficient to compute all values reported in this table: $\mathrm{P}(d=1) = 4\%$, $\mathrm{P}(x=1 \mid d=1) = 50\%$, $\mathrm{P}(x=1 \mid d=0) = 18.75\%$, $\mathrm{E}(y \mid d=1, x=1) = 3$, $\mathrm{E}(y \mid d=1, x=0) = 4$, $\mathrm{E}(y \mid d=0, x=1) = 0$, and $\mathrm{E}(y \mid d=0, x=0) = 5$. Hence, $\tau_1 = \mathrm{E}[y(1)-y(0) \mid x=1] = 3$ and $\tau_0 = \mathrm{E}[y(1)-y(0) \mid x=0] = -1$. The weights on τ_j, $j=0,1$, are *(i)* derived by Angrist (1998) for OLS, *(ii)* equal to $\mathrm{P}(x=j)$ for τ_{ATE}, *(iii)* equal to $\mathrm{P}(x=j \mid d=1)$ for τ_{ATT}, and *(iv)* equal to $\mathrm{P}(x=j \mid d=0)$ for τ_{ATC}. The weights on τ_{ATT} and τ_{ATC} are *(i)* derived in this paper for OLS, *(ii)* equal to $\mathrm{P}(d=1)$ and $\mathrm{P}(d=0)$, respectively, for τ_{ATE}, *(iii)* equal to 100% and 0%, respectively, for τ_{ATT}, and *(iv)* equal to 0% and 100%, respectively, for τ_{ATC}.

To see this fact, it is useful to focus on τ_{ATT} and τ_{ATC} instead. In practice, these parameters, together with τ_{ATE}, are usually more policy relevant than τ_1 and τ_0, especially in applications with many covariates. Let $\mathrm{E}(y \mid d=1, x=1) = 3$, $\mathrm{E}(y \mid d=1, x=0) = 4$, $\mathrm{E}(y \mid d=0, x=1) = 0$, and $\mathrm{E}(y \mid d=0, x=0) = 5$. What follows, $\tau_1 = 3$ and $\tau_0 = -1$, and hence the OLS estimand, τ, is equal to $48\% \cdot 3 + 52\% \cdot -1 = 0.92$. At the same time, however, it is straightforward to verify that $\tau_{ATE} = -0.2$, $\tau_{ATT} = 1$, and $\tau_{ATC} = -0.25$. See also Table 1 for more details. It turns out that the sign of the OLS estimand and the sign of τ_{ATE} are different; the average effect of treatment is negative but the OLS estimand is positive and relatively large in magnitude. Moreover, if we postulate that the OLS estimand is perhaps a convex combination of τ_{ATT} and τ_{ATC}, we can also write $0.92 = w_{ATT} \cdot 1 + (1 - w_{ATT}) \cdot -0.25$, where w_{ATT} is the OLS weight on τ_{ATT}. Solving for w_{ATT} yields $w_{ATT} = 93.6\%$. In other words, the OLS weight on the effect on the treated is very similar to the proportion of control units, 96%. Similarly, the OLS weight on the effect on the controls is surprisingly small and equal to 6.4%. Ordinary least squares approximately reverses the "natural" weights on τ_{ATT} and τ_{ATC}, equal to proportions of both groups. Table 1 gives a summary of this numerical example.

A crucial point of this paper is that this surprising result of "weight reversal" is not a coincidence—but instead a feature of OLS estimation. As I demonstrate below, the intuition from this simple example holds more generally.

Ordinary Least Squares

If $\mathrm{L}(\cdot \mid \cdot)$ denotes the linear projection, this paper is concerned with the interpretation of τ in the linear projection of y on d and X,

$$\mathrm{L}(y \mid 1, d, X) = \alpha + \tau d + X\beta, \tag{4}$$

when the population linear model is possibly incorrect. Before giving the main theoretical results of this paper, we need to introduce several definitions. In particular, let

$$\rho = \mathrm{P}(d = 1) \tag{5}$$

denote the unconditional probability of "treatment" and let

$$p(X) = \mathrm{L}(d \mid 1, X) = \alpha_p + X\beta_p \tag{6}$$

denote the "propensity score" from the linear probability model.[7] Note that $p(X)$ is the best linear approximation to the true propensity score. In principle, the specification in equations (4) and (6) can be arbitrarily flexible, so this linear approximation can be made very accurate; in fact, equation (4) can be thought of as partially linear, where we potentially include powers and cross-products of original control variables but exclude interactions between d and X.[8] It is also useful to note that the Frisch–Waugh theorem (Frisch and Waugh, 1933) implies that $\tau = \tau_a$, where τ_a is defined by

$$\mathrm{L}[y \mid 1, d, p(X)] = \alpha_a + \tau_a d + \gamma_a \cdot p(X). \tag{7}$$

In other words, the least squares estimand for the effect of d is the same whether we project y on d and covariates or on d and the "propensity score." This equivalence result will also hold if we replace population parameters with their sample analogues, as is the case elsewhere in this paper. While the equality of τ and τ_a might seem trivial, it is often overlooked and will be useful in proving my main result.

After defining $p(X)$, it is helpful to introduce two linear projections of y on $p(X)$,

[7] Note that this "propensity score" does not need to have any behavioral interpretation. For example, d can be an attribute, in the sense of Holland (1986), and therefore does not need to constitute a feasible "treatment" in any "ideal experiment" (Angrist and Pischke, 2009). Although it might be difficult, for example, to conceptualize the "propensity score" for gender or race, it does not matter for this definition.

[8] To be precise, I do not need to exclude interactions between d and X to prove my main result. This exclusion is only illustrative; it simplifies the interpretation of my result and it follows the standard practice of estimating the model in (1) without interactions.

separately for $d = 1$ and $d = 0$, namely

$$\mathrm{L}\left[y \mid 1, p\left(X\right)\right] = \alpha_1 + \gamma_1 \cdot p\left(X\right) \quad \text{if} \quad d = 1 \tag{8}$$

and also

$$\mathrm{L}\left[y \mid 1, p\left(X\right)\right] = \alpha_0 + \gamma_0 \cdot p\left(X\right) \quad \text{if} \quad d = 0. \tag{9}$$

Note that equations (6), (8), and (9) are definitional. I do not assume that these linear projections correspond to well-specified population models and I do not put any substantial restrictions on the underlying data-generating process. In fact, it is sufficient for my main result that the linear projections in (4), (6), (8), and (9) always exist and are unique.

Assumption 1 *(i)* $\mathrm{E}(y^2)$ *and* $\mathrm{E}(\|X\|^2)$ *are finite. (ii) The covariance matrices of* X *and* (d, X) *are nonsingular.*

Clearly, Assumption 1 guarantees the existence and uniqueness of the linear projections in (4) and (6). Additionally, Assumption 2 ensures that the linear projections in (8) and (9) also exist and are unique.

Assumption 2 $\mathrm{V}\left[p\left(X\right) \mid d = 1\right]$ *and* $\mathrm{V}\left[p\left(X\right) \mid d = 0\right]$ *are nonzero, where* $\mathrm{V}\left(\cdot \mid \cdot\right)$ *denotes the conditional variance (with respect to* $\mathrm{E}\left[p\left(X\right) \mid d = j\right]$*,* $j = 0, 1$*).*

Assumptions 1 and 2 are generally innocuous, although Assumption 2 rules out a small number of interesting applications, such as regression adjustments in Bernoulli trials and completely randomized experiments. In these cases, however, ordinary least squares is consistent for the average treatment effect under general conditions (see, *e.g.*, Imbens and Rubin, 2015). Moreover, as we will see later, Assumption 2 can be relaxed to permit either $\mathrm{V}\left[p\left(X\right) \mid d = 1\right]$ or $\mathrm{V}\left[p\left(X\right) \mid d = 0\right]$ to be zero.

The next step is to use the linear projections in (8) and (9) to define the average partial effect of d as

$$\tau_{APE} = \left(\alpha_1 - \alpha_0\right) + \left(\gamma_1 - \gamma_0\right) \cdot \mathrm{E}\left[p\left(X\right)\right] \tag{10}$$

as well as the average partial effect of d on group j ($j = 0, 1$) as

$$\tau_{APE|d=j} = \left(\alpha_1 - \alpha_0\right) + \left(\gamma_1 - \gamma_0\right) \cdot \mathrm{E}\left[p\left(X\right) \mid d = j\right]. \tag{11}$$

Because the linear projection passes through the point of means of all variables, namely $\mathrm{E}\left(y \mid d = 1\right) = \alpha_1 + \gamma_1 \cdot \mathrm{E}\left[p\left(X\right) \mid d = 1\right]$ and $\mathrm{E}\left(y \mid d = 0\right) = \alpha_0 + \gamma_0 \cdot \mathrm{E}\left[p\left(X\right) \mid d = 0\right]$, the average partial effects of d on both groups of interest can also be expressed as

$$\tau_{APE|d=1} = \mathrm{E}\left(y \mid d=1\right) - \left\{\alpha_0 + \gamma_0 \cdot \mathrm{E}\left[p\left(X\right) \mid d=1\right]\right\} \tag{12}$$

and also

$$\tau_{APE|d=0} = \left\{\alpha_1 + \gamma_1 \cdot \mathrm{E}\left[p\left(X\right) \mid d=0\right]\right\} - \mathrm{E}\left(y \mid d=0\right). \tag{13}$$

In other words, we only need the linear projection in (9), and not in (8), to define $\tau_{APE|d=1}$ because we already observe the mean treated outcome of the treated. Similarly, we need the linear projection in (8), but not in (9), to define $\tau_{APE|d=0}$, as the mean control outcome is observed for the control units. In either case we need to predict the mean outcomes of a given subpopulation in the alternative treatment regime. When both linear projections are well defined, $\tau_{APE|d=j}$ is also equivalent to the coefficient on d in the linear projection of y on d, $p\left(X\right)$, and $d \cdot \left\{p\left(X\right) - \mathrm{E}\left[p\left(X\right) \mid d=j\right]\right\}$.

Finally, if d is unconfounded (conditional on X) and $\mathrm{E}\left(d \mid X\right)$, $\mathrm{E}\left[y(1) \mid p\left(X\right)\right]$, and $\mathrm{E}\left[y(0) \mid p\left(X\right)\right]$ are linear, then τ_{APE}, $\tau_{APE|d=1}$, and $\tau_{APE|d=0}$ have a useful interpretation as τ_{ATE}, τ_{ATT}, and τ_{ATC}, respectively. It should be stressed, however, that the main result of this paper (Theorem 1) is more general and only requires Assumptions 1 and 2.

Theorem 1 (Weighted Average Interpretation of OLS) *Suppose that Assumptions 1 and 2 are satisfied. Then,*

$$\begin{aligned}\tau &= \frac{\rho \cdot \mathrm{V}\left[p\left(X\right) \mid d=1\right]}{\rho \cdot \mathrm{V}\left[p\left(X\right) \mid d=1\right] + (1-\rho) \cdot \mathrm{V}\left[p\left(X\right) \mid d=0\right]} \cdot \tau_{APE|d=0} \\ &+ \frac{(1-\rho) \cdot \mathrm{V}\left[p\left(X\right) \mid d=0\right]}{\rho \cdot \mathrm{V}\left[p\left(X\right) \mid d=1\right] + (1-\rho) \cdot \mathrm{V}\left[p\left(X\right) \mid d=0\right]} \cdot \tau_{APE|d=1}.\end{aligned}$$

Henceforth, to simplify notation, I will use w_0 to denote $\frac{\rho \cdot \mathrm{V}[p(X)|d=1]}{\rho \cdot \mathrm{V}[p(X)|d=1]+(1-\rho)\cdot \mathrm{V}[p(X)|d=0]}$ and w_1 to denote $\frac{(1-\rho) \cdot \mathrm{V}[p(X)|d=0]}{\rho \cdot \mathrm{V}[p(X)|d=1]+(1-\rho)\cdot \mathrm{V}[p(X)|d=0]}$.

A proof of Theorem 1 is provided in Appendix A. This theorem shows that τ, the ordinary least squares estimand, can be expressed as a convex combination of $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$. The definition of $\tau_{APE|d=j}$ makes it clear that the OLS estimand is identical to the outcome of a particular three-step procedure. In the first step, we obtain $p\left(X\right)$, *i.e.* the "propensity score." Next, in the second step, we obtain $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$, as in (11), from two linear projections of y on $p\left(X\right)$, separately for $d=1$ and $d=0$. Finally, in the third step, we calculate a weighted average of $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$. The weight which is placed by OLS on $\tau_{APE|d=1}$ is decreasing in $\frac{\mathrm{V}[p(X)|d=1]}{\mathrm{V}[p(X)|d=0]}$ and ρ and the weight which is placed on $\tau_{APE|d=0}$ is increasing in $\frac{\mathrm{V}[p(X)|d=1]}{\mathrm{V}[p(X)|d=0]}$ and ρ.[9]

The fact that Theorem 1 only requires the existence and uniqueness of several linear projections makes this result very general. On the other hand, one possible concern about this result might be that $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$ do not necessarily correspond to the causal objects of interest, unless the additional linearity restrictions are satisfied. I address this issue in detail below. In general, it is possible to decompose the difference between τ and each causal object of interest into components attributable to *(i)* these linearity restrictions and *(ii)* implicit weights on $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$.

The weighting scheme in Theorem 1 might be seen as surprising: the more units belong to group j, the less weight is placed on $\tau_{APE|d=j}$, *i.e.* the effect *on this group*. To aid intuition, recall that an important motivation for using ordinary least squares to estimate the model in (1) is that the linear projection of y on d and X provides the best linear predictor of y given d and X (see, *e.g.*, Angrist and Pischke, 2009). However, if our goal is to conduct causal inference, then this is not, in fact, a good reason to use this method. Ordinary least squares is "best" in predicting actual outcomes, while causal inference is about predicting missing outcomes, defined as $y_m = y(1) \cdot (1-d) + y(0) \cdot d$. In other words, the OLS weights are optimal for predicting "what is." Instead, we are interested in predicting "what would be" if treatment were assigned differently.

Intuition suggests that if our goal were in predicting "what is" and, without loss of generality, group one were substantially larger than group zero, we would like to place a large weight on the coefficients of group one (α_1 and γ_1), because these coefficients would be used to predict actual outcomes of this group. Clearly, as noted by Deaton (1997) and Solon *et al.* (2015), the OLS weights are consistent with this idea. However, if our goal is to predict missing outcomes, we need to place a large weight on the coefficients of group zero, because these coefficients are used to predict counterfactuals for group one. It is also useful to note that the (infeasible) linear projection of the missing outcome, y_m, on d and X would solve our problem of "weight reversal." The weights on $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$ would still be different than ρ and $1-\rho$ if the conditional variances of $p(X)$ were different in the two groups; but, at least, the weight on $\tau_{APE|d=1}$ ($\tau_{APE|d=0}$) would be increasing (decreasing) in ρ.[10]

[9] In fact, a more formal treatment of the relationship between ρ and w_1 (w_0) is necessary to determine that this relationship is indeed always negative (positive). A proof of this proposition, which additionally assumes that the population model for d is linear in X, is provided in Appendix B. In this proof, I study the derivatives of w_1 and w_0 with respect to the intercept of the propensity score model. Clearly, such an intercept shift is equivalent to a change in ρ.

[10] To aid understanding of the OLS weights, it is also useful to consider partial residualization that is

There are several interesting corollaries of Theorem 1. Similar to the discussion above, Corollary 1 clarifies the causal interpretability of the OLS estimand.

Corollary 1 (Causal Interpretation of OLS) *Suppose that d is unconfounded conditional on covariates, X. Also, suppose that the population models for d and both of $y(1)$ and $y(0)$ are linear in X and $p(X)$, respectively. Then, Theorem 1 implies that*

$$\tau \quad = \quad w_0 \cdot \tau_{ATC} + w_1 \cdot \tau_{ATT}.$$

In other words, if d is unconfounded conditional on X and the population models for d and both of $y(1)$ and $y(0)$ are linear in X and $p(X)$, respectively, the weighting scheme from Theorem 1 will apply to τ_{ATT} and τ_{ATC}.[11] Indeed, the weight which is placed on τ_{ATT} is decreasing in ρ and the weight which is placed on τ_{ATC} is increasing in ρ.

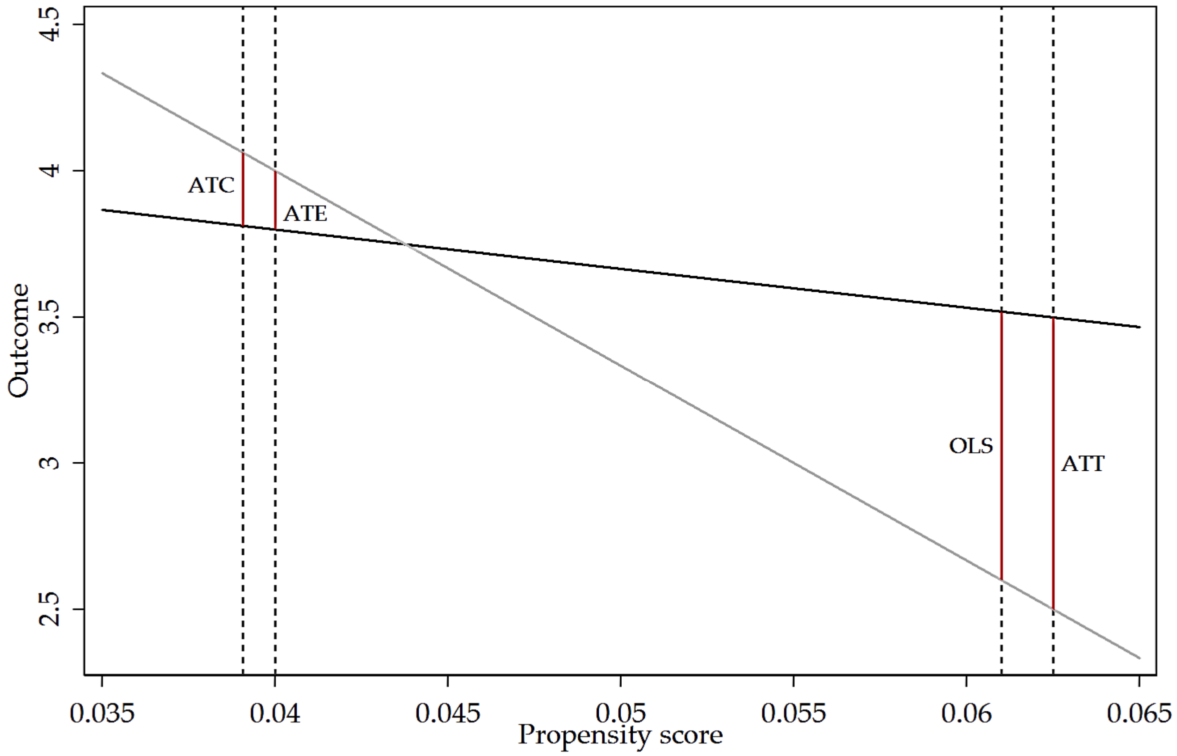
While the linearity assumptions for $\mathrm{E}(d \mid X)$, $\mathrm{E}[y(1) \mid p(X)]$, and $\mathrm{E}[y(0) \mid p(X)]$ are restrictive, they are not unusual in the recent literature. The linearity of $\mathrm{E}(d \mid X)$ is assumed in Rhodes (2010), Aronow and Samii (2016), and Abadie, Athey, Imbens, and Wooldridge (2017). It is also implicit in the analysis of saturated models by Angrist (1998) and Humphreys (2009). The linearity assumption for $\mathrm{E}[y(1) \mid p(X)]$ and $\mathrm{E}[y(0) \mid p(X)]$ is also not new. See Wooldridge (2010) for a textbook discussion. A similar assumption is also used by Brinch, Mogstad, and Wiswall (2017) to identify marginal treatment effects with a binary instrument. In general, the (non)linearity of potential outcomes with respect to the propensity score figures prominently in the literature on marginal treatment effects (see, *e.g.*, Heckman and Vytlacil, 2005; Moffitt, 2008).

A graphical illustration of Theorem 1 (and Corollary 1) is provided in Figure 1, which gives two linear projections of y on $p(X)$, separately for each treatment status. This figure corresponds to the numerical example in Table 1 above, and hence $\tau = 0.92$, $\tau_{ATE} = -0.2$, $\tau_{ATT} = 1$, and $\tau_{ATC} = -0.25$. The average treatment effect is equal to the distance between the two linear projections at the mean of $p(X)$. Similarly, τ_{ATT} and τ_{ATC} are equal to the same distance, evaluated at the group-specific means. Because the linearity assumptions for $\mathrm{E}(d \mid X)$, $\mathrm{E}[y(1) \mid p(X)]$, and $\mathrm{E}[y(0) \mid p(X)]$ are in fact satisfied in this example, each of these objects has a causal interpretation. Also, since there are relatively few treated units, τ_{ATE} is much closer to τ_{ATC} than to τ_{ATT}. However, the opposite is true

implicit in least squares estimation. This idea is pursued further in Appendix C.

[11] Usually, we would also require that there is complete overlap in the support of the distributions of X among treated and control units, but instead we implicitly postulate a (globally) linear relationship between potential outcomes and $p(X)$. This allows us to extrapolate even in the absence of overlap.

Figure 1: Graphical Illustration of Theorem 1 (Corollary 1)



Notes: This figure provides a graphical illustration of Theorem 1 (and Corollary 1). The example presented in this figure corresponds to the numerical example presented in Table 1. Linear projections of y on $p\left(X\right)$ are represented by solid lines, black for treated units and gray for control units. Parameters of interest are represented by solid vertical red segments that measure the distance between both linear projections at specific values of $p\left(X\right)$, represented by dashed vertical lines.

for the OLS estimand, which is equal to the distance between the two linear projections at $w_1 \cdot \mathrm{E}\left[p\left(X\right) \mid d=1\right] + w_0 \cdot \mathrm{E}\left[p\left(X\right) \mid d=0\right] \simeq 93.6\% \cdot 0.0625 + 6.4\% \cdot 0.039 = 0.061$.

Importantly, even if we were unwilling to maintain unconfoundedness or the linearity assumptions for $\mathrm{E}\left(d \mid X\right)$, $\mathrm{E}\left[y(1) \mid p\left(X\right)\right]$, and $\mathrm{E}\left[y(0) \mid p\left(X\right)\right]$ in a different context, our graphical illustration would remain unchanged; only the labeling of some of our parameters of interest would be affected. According to Theorem 1, the OLS estimand will *always* remain equal to the distance between the two linear projections at $w_1 \cdot \mathrm{E}\left[p\left(X\right) \mid d=1\right] + w_0 \cdot \mathrm{E}\left[p\left(X\right) \mid d=0\right]$. The only difference is that causal statements would be inappropriate without the additional linearity assumptions in Corollary 1, and the objects previously referred to as average treatment effects would need to be termed differently, as in Theorem 1.

In fact, according to Corollary 2, it is possible to decompose the difference between τ, the ordinary least squares estimand, and τ_{ATE}, the average treatment effect, into components attributable to *(i)* these linearity assumptions ("bias from linearity") and *(ii)* implicit weights on heterogeneous treatment effects ("bias from heterogeneity").

Corollary 2 *Theorem 1 implies that (i)*

$$\tau - \tau_{ATE} = \underbrace{w_0 \cdot \left(\tau_{APE|d=0} - \tau_{ATC}\right) + w_1 \cdot \left(\tau_{APE|d=1} - \tau_{ATT}\right)}_{\textit{bias from linearity}} + \underbrace{\delta \cdot \left(\tau_{ATC} - \tau_{ATT}\right)}_{\textit{bias from heterogeneity}},$$

where

$$\delta = \frac{\rho^2 \cdot \mathrm{V}\left[p\left(X\right) \mid d=1\right] - \left(1-\rho\right)^2 \cdot \mathrm{V}\left[p\left(X\right) \mid d=0\right]}{\rho \cdot \mathrm{V}\left[p\left(X\right) \mid d=1\right] + \left(1-\rho\right) \cdot \mathrm{V}\left[p\left(X\right) \mid d=0\right]} = \rho - w_1 = w_0 - \left(1-\rho\right),$$

and also that (ii)

$$\tau - \tau_{APE} = \delta \cdot \left(\tau_{APE|d=0} - \tau_{APE|d=1}\right).$$

Suppose that $\tau_{APE|d=1} \neq \tau_{APE|d=0}$. Then, Theorem 1 implies that (iii)

$$\tau = \tau_{APE} \qquad \textit{if and only if} \qquad \frac{\mathrm{V}\left[p\left(X\right) \mid d=1\right]}{\mathrm{V}\left[p\left(X\right) \mid d=0\right]} = \left(\frac{1-\rho}{\rho}\right)^2.$$

Corollary 2 establishes several important implications of Theorem 1. The starting point is a bias formula for OLS estimation of the model in (1) where we control for covariates but do not interact these covariates with d. It is clear that there would be no bias in estimating τ_{ATE} if average treatment effects were equal for both groups of interest, $\tau_{ATT} = \tau_{ATC}$, given that correct functional forms have been specified, namely $\tau_{ATT} = \tau_{APE|d=1}$ and $\tau_{ATC} = \tau_{APE|d=0}$; see result *(i)* above. At the same time, if treatment effects are heterogeneous, then correct functional forms will not solve the problem, except in some special cases. In fact, if we ignore the linearity assumptions and focus on estimating τ_{APE}, then we can greatly simplify our bias formula; see result *(ii)*.

Indeed, the difference between τ, the ordinary least squares estimand, and τ_{APE}, the average partial effect, depends only on a particular measure of heterogeneity, *i.e.* the difference between $\tau_{APE|d=0}$ and $\tau_{APE|d=1}$, and the parameter δ. A few comments on δ are in order. First, it is easy to verify that $-1 < \delta < 1$. What follows, $|\delta|$ has an intuitive interpretation as the percentage of $\operatorname{sgn}\left(\delta\right) \cdot \left(\tau_{APE|d=0} - \tau_{APE|d=1}\right)$ which is equal to the difference between τ and τ_{APE}. If δ is close to zero, then this difference will be small, unless the amount of heterogeneity is extreme. On the other hand, if δ is far from zero, then neglecting heterogeneity will be problematic, unless treatment effects are, in fact, homogeneous. Second, the sign of δ will be informative about the direction of the bias if we have any information or prior beliefs about the sign of $\tau_{APE|d=0} - \tau_{APE|d=1}$. Third, δ is a function of observable quantities and can be easily estimated. In fact, estimating δ requires no data on outcomes, and hence it can serve as a diagnostic before any further analysis has been carried out. Testing $\delta = 0$ is also straightforward. Finally, the calculation of δ is further simplified under the restriction that $\mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right]$. If we use δ^* to denote the value of this parameter in such a special case, it turns out that $\delta^* = 2\rho - 1$. Apparently, in this setting, the knowledge of δ only requires information on ρ, the proportion of units with $d = 1$. For example, if $\rho = 5\%$, then $\delta^* = -90\%$; if $\rho = 60\%$, then $\delta^* = 20\%$. Of course, the special case where $\mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right]$ is hardly to be expected in practice. Still, $\delta^* = 2\rho - 1$ can potentially serve as a rule of thumb. For example, if one group is much larger than another, and hence δ^* is close to one in absolute value, the bias from heterogeneity can easily become an issue. On the other hand, if the proportions of units with $d = 1$ and $d = 0$ are roughly equal, we will expect τ to approximately recover τ_{APE}, as $\rho = 50\%$ implies that $\delta^* = 0$.[12]

This rule of thumb—that is, that OLS estimation of the model in (1) will be approximately consistent for τ_{APE} whenever both groups of interest are of similar size—can also be seen from a slightly different perspective. We can start with noting that the average partial effect of d can be written as

$$\tau_{APE} = \rho \cdot \tau_{APE|d=1} + (1 - \rho) \cdot \tau_{APE|d=0}. \tag{14}$$

Then, Corollary 3 provides an already familiar condition under which ordinary least squares reverses these "natural" weights on $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$.

Corollary 3 *Suppose that* $\mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right]$. *Then, Theorem 1 implies that*

$$\tau = \rho \cdot \tau_{APE|d=0} + (1 - \rho) \cdot \tau_{APE|d=1}.$$

Precisely, if the variance of the "propensity score" is equal in both groups of interest, then the OLS estimand is equal to a weighted average of both group-specific average

[12] Corollary 2 also provides a general condition under which $\tau = \tau_{APE}$; see result *(iii)*. Namely, the OLS estimand is equal to the average partial effect only in a special case, where the ratio of the conditional variances of the "propensity score" is equal to the square of the reversed ratio of population proportions of both groups of interest. In general, there can be little reason to expect this condition to hold.

partial effects, with "reversed" weights attached to these effects. Namely, the proportion of units with $d = 1$ is used to weight the average partial effect of d on group zero and the proportion of units with $d = 0$ is used to weight the average partial effect of d on group one. Therefore, there is only one situation in which Corollary 3 allows the OLS estimand to be equal to the average partial effect of d, and this occurs, as we have already seen, whenever not only $\mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right]$ but also $\rho = 1 - \rho = 50\%$.

Corollary 3 provides the foundation for a yet another rule of thumb: if the vast majority of units belong to group j, then the OLS estimand will be approximately equal to the effect *on the other group*. If ρ is close to zero (one), then τ is close to $\tau_{APE|d=1}$ ($\tau_{APE|d=0}$). In other words, if we estimate the model in (1) using OLS and there are relatively very few (many) treated units, then we might be willing to interpret our estimate as that of the effect on the treated (controls). This approximation is never exactly correct, as ρ is strictly between zero and one, but it might work well enough in practice. This relationship between ρ and τ can also be seen by decomposing the differences between τ and τ_{ATT} as well as between τ and $\tau_{APE|d=1}$, similar to Corollary 2.

Corollary 4 *Theorem 1 implies that (i)*

$$\begin{aligned}\tau - \tau_{ATT} &= \underbrace{w_0 \cdot \left(\tau_{APE|d=0} - \tau_{ATC}\right) + w_1 \cdot \left(\tau_{APE|d=1} - \tau_{ATT}\right)}_{\textit{bias from linearity}} \\ &+ \underbrace{w_0 \cdot \left(\tau_{ATC} - \tau_{ATT}\right)}_{\textit{bias from heterogeneity}},\end{aligned}$$

and also that (ii)

$$\tau - \tau_{APE|d=1} = w_0 \cdot \left(\tau_{APE|d=0} - \tau_{APE|d=1}\right).$$

Indeed, according to Corollary 4, the bias from heterogeneity—when we turn our attention to the effect on the treated—depends only on w_0 as well as on the difference between $\tau_{APE|d=0}$ and $\tau_{APE|d=1}$ (or τ_{ATC} and τ_{ATT}). This bias is also increasing in w_0 while w_0, in turn, is increasing in ρ. What follows, if there are few treated units, w_0 is small and bias is small; if there are many treated units, the bias is relatively large. Also, w_0 can be interpreted, like $|\delta|$, as the percentage of our measure of heterogeneity, $\tau_{APE|d=0} - \tau_{APE|d=1}$, which contributes to bias. Like δ, w_0 is also easily estimable and can be used as a simple diagnostic. Unlike in the case of δ, it makes no sense, however, to test $w_0 = 0$, as it is always the case that $0 < w_0 < 1$. In other words, under Assumptions 1 and 2, and with heterogeneous treatment effects, there is always nonzero bias from OLS estimation of the model in (1), with the effect on the treated as the parameter of interest. This does not have to be the case for the average partial effect, although, as we have seen, nonzero bias should also be expected (cf. footnote 12). Finally, if we use w_0^* to denote the value of w_0 that is consistent with the restriction that $\mathrm{V}\left[p\left(X\right) \mid d=1\right] = \mathrm{V}\left[p\left(X\right) \mid d=0\right]$, it turns out that this rule of thumb is simply $w_0^* = \rho$. Again, if there are very few treated units, we can expect to approximately recover the effect on the treated; if there are relatively many treated units, we might not recover either of τ_{APE} and $\tau_{APE|d=1}$.

Related Literature

It is useful to discuss the relationship between Theorem 1 and the previous results in Angrist (1998) and Humphreys (2009). On the one hand, there is limited overlap between my main result and these previous contributions, because they both restrict their attention to saturated models with discrete covariates, in which the estimating equation includes a binary variable for each combination of covariate values ("stratum"). In this paper I provide a more general result which is not restricted to saturated models or discrete covariates. On the other hand, some connections between these contributions can nevertheless be made. First, note that the baseline result in Angrist (1998) is derived for a model with only two strata. Appendix D demonstrates that this result follows from a special case of Theorem 1, in which X is a single binary variable. Hence, Theorem 1 is more general than the baseline result in Angrist (1998), and it also provides a substantially different interpretation of the ordinary least squares estimand. Aside from the result for a model with two strata, Angrist (1998) also provides a more general representation of τ_n in

$$\mathrm{L}\left(y \mid d, x_1, \ldots, x_S\right) = \tau_n d + \sum_{s=1}^{S} \beta_{n,s} x_s, \tag{15}$$

where $x_1, \ldots, x_S$ are stratum indicators. More precisely, Angrist (1998, fn. 11) demonstrates that

$$\tau_n = \sum_{s=1}^{S} \frac{\mathrm{P}\left(x_s = 1\right) \cdot \mathrm{P}\left(d=1 \mid x_s = 1\right) \cdot \mathrm{P}\left(d=0 \mid x_s=1\right)}{\sum_{t=1}^{S} \mathrm{P}\left(x_t = 1\right) \cdot \mathrm{P}\left(d=1 \mid x_t = 1\right) \cdot \mathrm{P}\left(d=0 \mid x_t=1\right)} \cdot \tau_s, \tag{16}$$

where $\tau_s = \mathrm{E}\left(y \mid d=1, x_s=1\right) - \mathrm{E}\left(y \mid d=0, x_s=1\right)$. This representation of the OLS estimand is clearly different from Theorem 1 and Corollary 1, and each result offers certain advantages. On the one hand, unlike Corollary 1, Angrist (1998) does not have to restrict the relationship between τ_s and $\mathrm{P}\left(d=1 \mid x_s=1\right)$ to be linear. On the other hand, unlike this paper, Angrist (1998) restricts his attention to saturated models with discrete covariates, which are clearly different from models that are typically estimated in practice. Finally, Theorem 1 and Corollary 1 make it arguably easier to identify whether in a given empirical application the ordinary least squares estimand will be close to any of the parameters of interest (cf. Corollaries 2 to 4). In particular, Angrist (1998) does not recover a pattern of "weight reversal," which is discussed in detail in this paper.

Second, note that Humphreys (2009) does not, in fact, derive a new representation of τ_n, but instead provides a more detailed analysis of the result in Angrist (1998). In particular, Humphreys (2009) notes that τ_n, as represented in (16), can take any value between $\min(\tau_s)$ and $\max(\tau_s)$. Then, he demonstrates that τ_n is also bounded by τ_{ATT} and τ_{ATC} if we restrict the relationship between τ_s and $\mathrm{P}\left(d = 1 \mid x_s = 1\right)$ to be monotonic. According to Corollary 1, τ is a convex combination of τ_{ATT} and τ_{ATC} if, among other things, the population models for $y(1)$ and $y(0)$ are linear in $p\left(X\right)$, which also implies a linear relationship between τ_s and $\mathrm{P}\left(d = 1 \mid x_s = 1\right)$ when the model for y is saturated. Of course, this linearity assumption in Corollary 1 is stronger than the monotonicity assumption in Humphreys (2009). However, in return, we are able to derive a closed-form expression for τ in terms of τ_{ATT} and τ_{ATC}. Also, like Angrist (1998), Humphreys (2009) restricts his attention to saturated models with discrete covariates, which is a major limitation.

Estimating Heterogeneous Effects

There are several constructive solutions to the problem described in this paper. First, it is sufficient to interact the variable of interest with other covariates, and then calculate its average partial effect on a given group (similar to equations (10) and (11)). This leads to an estimator which is sometimes referred to as "Oaxaca–Blinder" (Kline, 2011, 2014), "regression adjustment" (Wooldridge, 2010), "flexible OLS" (Khwaja, Picone, Salm, and Trogdon, 2011), or even "regression" (Imbens and Wooldridge, 2009). Second, if one is not comfortable with the linear approximation to the conditional mean, it is possible to use any of the standard semi- and nonparametric estimators for average treatment effects, such as inverse probability weighting, matching, and other methods based on the propensity score (for a review, see Imbens and Wooldridge, 2009). Third, it might also help to estimate a model with homogeneous effects using weighted least squares. In particular, we might use weights to estimate a model with d and $p\left(X\right)$ as the only independent variables. In this case we would like to characterize a set of weights, w, such that τ_w in

$$\mathrm{L}\left(\sqrt{w} \cdot y \mid \sqrt{w}, \sqrt{w} \cdot d, \sqrt{w} \cdot p\left(X\right)\right) = \alpha_w \cdot \sqrt{w} + \tau_w \cdot \sqrt{w} \cdot d + \gamma_w \cdot \sqrt{w} \cdot p\left(X\right) \qquad (17)$$

has a useful interpretation. An appropriate set of weights is provided in Theorem 2.

Theorem 2 (Weighted Least Squares Correction) *Suppose that Assumptions 1 and 2 are satisfied. Also, $w = \frac{1-\rho}{w_0} \cdot d + \frac{\rho}{w_1} \cdot (1-d)$. Then,*

$$\tau_w = \tau_{APE}.$$

The proof of Theorem 2 follows directly from the proof of Theorem 1. Indeed, the average partial effect of d can be recovered from a weighted least squares procedure, with weights of $\frac{1-\rho}{w_0}$ for units with $d = 1$ and weights of $\frac{\rho}{w_1}$ for units with $d = 0$. These weights consist of two parts: either $\frac{1}{w_1}$ or $\frac{1}{w_0}$; and either ρ or $1-\rho$. The role of the first part is always to undo the OLS weights (w_1 and w_0 in Theorem 1); the role of the second part is to impose the correct weights of ρ on $\tau_{APE|d=1}$ and $1-\rho$ on $\tau_{APE|d=0}$. Finally, it is useful to note that there is no similar procedure to recover $\tau_{APE|d=1}$ or $\tau_{APE|d=0}$; both of these objects, however, are easily obtained from equation (11).

Interestingly, the structure of the weighted least squares procedure in Theorem 2 resembles the "tyranny of the minority" estimator in Lin (2013). This method uses weights of $\frac{1-\rho}{\rho}$ for units with $d = 1$ and weights of $\frac{\rho}{1-\rho}$ for units with $d = 0$; it also controls for X instead of $p(X)$. It is important to note, however, that this method is designed to solve a different problem than Theorem 2. In particular, Freedman (2008b,a) demonstrates that regression adjustments to experimental data can lead to a loss in precision. On the other hand, Lin (2013) shows that this is no longer possible if we additionally interact d with X. Then, he derives the "tyranny of the minority" estimator as an alternative least squares procedure, based on a single conditional mean, which does not suffer from this loss in precision. In the context of observational data, however, this estimator is consistent for the average partial effect of d only in a special case, namely under the restriction in Corollary 3, $\mathrm{V}[p(X) \mid d = 1] = \mathrm{V}[p(X) \mid d = 0]$.

Two-Stage Least Squares

Finally, I demonstrate that a result similar to Theorem 1 carries over to two-stage least squares. As before, let Z denote the row vector of instrumental variables for d, $(z_1, \ldots, z_L)$. It is helpful to start with introducing an additional linear projection, namely

$$d(X, Z) = \mathrm{L}(d \mid 1, X, Z) = \alpha_f + X\beta_f + Z\zeta_f, \tag{18}$$

where $d(X,Z)$ are the fitted values from a linear first stage. It is then well known that the two-stage least squares estimand is equal to τ_{ts} in

$$\mathrm{L}\left[y \mid 1, d(X,Z), X\right] = \alpha_{ts} + \tau_{ts} \cdot d(X,Z) + X\beta_{ts}. \tag{19}$$

However, it is not possible to apply Theorem 1 to τ_{ts} because $d(X,Z)$ is not binary. Instead, it is sufficient to note that $\tau_{ts} = \tau_{cf}$ if

$$\mathrm{L}\left(y \mid 1, d, X, r\right) = \alpha_{cf} + \tau_{cf} \cdot d + X\beta_{cf} + \eta_{cf} r, \tag{20}$$

where

$$r = d - d(X,Z) \tag{21}$$

are the errors from a linear first stage. They serve as a control function, *i.e.* a variable which renders d appropriately exogenous (see, *e.g.*, Wooldridge, 2015).[13]

Because d is binary and $\tau_{ts} = \tau_{cf}$, we can provide a new interpretation of the 2SLS estimand by applying Theorem 1 to τ_{cf}. To accomplish this, we need to redefine several objects from before. In particular, let the "propensity score" be denoted by

$$p(X,r) = \mathrm{L}\left(d \mid 1, X, r\right) = \alpha_{s,ts} + X\beta_{s,ts} + \eta_{s,ts} r. \tag{22}$$

Similarly, we need two auxiliary linear projections of y on $p(X,r)$, separately for $d = 1$ and $d = 0$, namely

$$\mathrm{L}\left[y \mid 1, p(X,r)\right] = \alpha_{1,ts} + \gamma_{1,ts} \cdot p(X,r) \quad \text{if} \quad d = 1, \tag{23}$$

and also

$$\mathrm{L}\left[y \mid 1, p(X,r)\right] = \alpha_{0,ts} + \gamma_{0,ts} \cdot p(X,r) \quad \text{if} \quad d = 0. \tag{24}$$

As before, equations (22) to (24) are definitional; even though equations (23) and (24) are similar to potential outcome models in Brinch *et al.* (2017), they do not have to represent well-specified conditional expectations. We do require, however, that the linear projections listed above exist and are unique. This is guaranteed by Assumptions 3 and 4.

Assumption 3 *(i)* $\mathrm{E}(y^2)$*,* $\mathrm{E}(\|X\|^2)$*, and* $\mathrm{E}(\|Z\|^2)$ *are finite. (ii) The covariance matrices of* (X,Z)*,* (d,X,r)*, and* (X,r) *are nonsingular.*

[13] This control function representation of two-stage least squares is often attributed to Hausman (1978), although it seems to go back to earlier literature. For example, Blundell and Matzkin (2014) note that a similar result is discussed in Telser (1964). See also Heckman (1978) and Kline and Walters (2018).

Assumption 4 $\mathrm{V}\left[p\left(X,r\right) \mid d=1\right]$ *and* $\mathrm{V}\left[p\left(X,r\right) \mid d=0\right]$ *are nonzero.*

Because the linear projections in (23) and (24) are well defined by assumption, they can be used to define the average partial effect of d as

$$\tau_{APE,ts} = \left(\alpha_{1,ts} - \alpha_{0,ts}\right) + \left(\gamma_{1,ts} - \gamma_{0,ts}\right) \cdot \mathrm{E}\left[p\left(X,r\right)\right] \tag{25}$$

as well as the average partial effect of d on group j ($j = 0, 1$) as

$$\tau_{APE,ts|d=j} = \left(\alpha_{1,ts} - \alpha_{0,ts}\right) + \left(\gamma_{1,ts} - \gamma_{0,ts}\right) \cdot \mathrm{E}\left[p\left(X,r\right) \mid d=j\right]. \tag{26}$$

These objects are used in Theorem 3 to provide a general weighted average representation of the two-stage least squares estimand.

Theorem 3 (Weighted Average Interpretation of 2SLS) *Suppose that Assumptions 3 and 4 are satisfied. Then,*

$$\begin{aligned} \tau_{ts} &= \frac{\rho \cdot \mathrm{V}\left[p\left(X,r\right) \mid d=1\right]}{\rho \cdot \mathrm{V}\left[p\left(X,r\right) \mid d=1\right] + \left(1-\rho\right) \cdot \mathrm{V}\left[p\left(X,r\right) \mid d=0\right]} \cdot \tau_{APE,ts|d=0} \\ &+ \frac{\left(1-\rho\right) \cdot \mathrm{V}\left[p\left(X,r\right) \mid d=0\right]}{\rho \cdot \mathrm{V}\left[p\left(X,r\right) \mid d=1\right] + \left(1-\rho\right) \cdot \mathrm{V}\left[p\left(X,r\right) \mid d=0\right]} \cdot \tau_{APE,ts|d=1}. \end{aligned}$$

Henceforth, to simplify notation, I will use $w_{0,ts}$ *to denote* $\frac{\rho \cdot \mathrm{V}[p(X,r)|d=1]}{\rho \cdot \mathrm{V}[p(X,r)|d=1] + (1-\rho) \cdot \mathrm{V}[p(X,r)|d=0]}$ *and* $w_{1,ts}$ *to denote* $\frac{(1-\rho) \cdot \mathrm{V}[p(X,r)|d=0]}{\rho \cdot \mathrm{V}[p(X,r)|d=1] + (1-\rho) \cdot \mathrm{V}[p(X,r)|d=0]}$.

The proof of Theorem 3 is analogous to that of Theorem 1, after noticing that $\tau_{ts} = \tau_{cf}$. Theorem 3 makes it clear that the 2SLS estimand is identical to the outcome of a procedure that is, not surprisingly, very similar to OLS. In the first step, we project d on all exogenous variables ("first stage"). In the second step, we obtain $p\left(X,r\right)$, *i.e.* the "propensity score" from a linear projection of d on X and the first-stage errors. In the third step, we obtain $\tau_{APE,ts|d=1}$ and $\tau_{APE,ts|d=0}$ from two linear projections of y on $p\left(X,r\right)$, separately for $d=1$ and $d=0$. In the fourth step, we calculate a weighted average of $\tau_{APE,ts|d=1}$ and $\tau_{APE,ts|d=0}$. The weight which is placed by two-stage least squares on $\tau_{APE,ts|d=1}$ is decreasing in $\frac{\mathrm{V}[p(X,r)|d=1]}{\mathrm{V}[p(X,r)|d=0]}$ and ρ and the weight which is placed on $\tau_{APE,ts|d=0}$ is increasing in $\frac{\mathrm{V}[p(X,r)|d=1]}{\mathrm{V}[p(X,r)|d=0]}$ and ρ. Clearly, Theorem 3 also has several implications that are analogous to Corollaries 1 to 4 and Theorem 2. For conciseness they are not restated here.[14]

It is also useful to explain the differences between this implicit estimation procedure and the Heckman's two-step estimation procedure for the Gaussian switching regime model (Heckman, 1976, 1979). As noted by Heckman, Tobias, and Vytlacil (2001, 2003) and Wooldridge (2015), this latter procedure can be used for estimation of τ_{ATE}, τ_{ATT}, and τ_{ATC} using instrumental variables. The first difference between 2SLS and the Heckman's procedure is in the computation of the first-stage residuals. The Heckman's procedure uses a probit first stage and generalized residuals, while 2SLS is implicitly based on a linear first stage and OLS residuals.[15] Second, in the next step, the Heckman's procedure amounts to regressing y on the vector of other covariates and the generalized residuals, separately for each treatment status, while 2SLS is implicitly based on similar regressions of y on the "propensity score," which in turn depends on other covariates and the first-stage residuals. Third, and most importantly, the Heckman's procedure allows us to calculate appropriate averages of the difference in fitted values from the previous step—and this allows us to estimate each of τ_{ATE}, τ_{ATT}, and τ_{ATC}—while two-stage least squares is equal to a specific weighted average of the estimated effects on the treated and controls, where the weights are inversely related to the proportion of each group.

Two further remarks are in order. First, Heckman and Vytlacil (2005) conclude that the IV estimand is equal to the average treatment effect when potential outcomes are linear in the propensity score. Since the analysis in Heckman and Vytlacil (2005) is made conditional on X, Theorem 3 makes it clear that this conclusion does not generally extend to unconditional estimands. Second, in a setting without additional covariates, Kline and Walters (2018) demonstrate that the IV estimator is algebraically equivalent to a number of control function estimators of the local average treatment effect, including an estimator based on the Heckman's two-step procedure. This equivalence disappears, however, when covariates are introduced into the model. An implication of Theorem 3 is that even in a setting with covariates, the 2SLS estimand has an implicit structure that is similar to control function procedures; it is also different, however, from any of the standard parameters of interest. A more detailed comparison between the IV estimand and the local average treatment effect is relegated toward the end of the paper.

[14] For later reference, let the 2SLS analogue of δ be denoted by $\delta_{ts} = \frac{\rho^2 \cdot \mathrm{V}[p(X,r)|d=1] - (1-\rho)^2 \cdot \mathrm{V}[p(X,r)|d=0]}{\rho \cdot \mathrm{V}[p(X,r)|d=1] + (1-\rho) \cdot \mathrm{V}[p(X,r)|d=0]} = \rho - w_{1,ts} = w_{0,ts} - (1-\rho)$. Interestingly, $w^*_{0,ts} = \rho$ and $\delta^*_{ts} = 2\rho - 1$. In other words, the "rule of thumb" values of these diagnostics, now obtained under the restriction that $\mathrm{V}\left[p\left(X,r\right) \mid d=1\right] = \mathrm{V}\left[p\left(X,r\right) \mid d=0\right]$, are identical for OLS and 2SLS.

[15] As noted by Olsen (1980), replacing a probit selection equation (first stage) with a linear selection equation (first stage) is equivalent to assuming that the error term in the selection (treatment) equation is uniformly distributed. Otherwise this procedure is analogous to Heckman (1976, 1979).

3 Empirical Applications

This section provides several empirical illustrations of Theorems 1 and 3 as well as their various corollaries.[16] I begin with outlining four scenarios that applied researchers are likely to face, and then provide an empirical application for each of them. I assume that the researcher wishes to estimate the model in (1) using OLS or 2SLS but is concerned about treatment effect heterogeneity; she might be interested in ATE, ATT, or both.

It is important to note that throughout this section τ_{APE}, $\tau_{APE|d=1}$, and $\tau_{APE|d=0}$ are implicitly treated as our parameters of interest. Although this might be restrictive, I also demonstrate that in all my empirical applications sample analogues of these parameters, reported in the body of the paper, are very similar to other estimates of ATE, ATT, and ATC, reported in Appendix F. In other words, this section focuses on the "bias from heterogeneity" and largely ignores the "bias from linearity," but my results also suggest that the latter source of bias is not particularly important in these applications. Of course, in other empirical contexts, the bias from linearity might also be substantial.

The first of four scenarios is that the researcher is only interested in one of these parameters and the diagnostic methods in Corollaries 2 and 4 indicate that this parameter should approximately be recovered by OLS or 2SLS, even if treatment effects are heterogeneous. The analysis of the effects of the National Supported Work (NSW) program in Angrist and Pischke (2009) is consistent with this scenario. In this application, the proportion of treated units is so small that OLS estimates are very similar to those of the effect on the treated, even though the amount of heterogeneity is quite substantial.

The second scenario is that the diagnostic methods indicate that OLS or 2SLS estimates might be quite different from those of our parameter or parameters of interest, but, in fact, they are similar, as treatment effects are also relatively homogeneous. To provide an empirical illustration of this scenario, I replicate the analysis of anti-Semitic violence in Voigtländer and Voth (2012). The effects of medieval pogroms on twentieth-century anti-Semitism are sufficiently homogeneous to outweigh the fact that my diagnostics would otherwise suggest caution.

In the third scenario, the researcher is potentially interested in both ATE and ATT, and the diagnostics indicate that one of these parameters—but of course not both—should be recovered by OLS or 2SLS even in the presence of treatment effect heterogeneity. The analysis of the effects of Catholic schooling on twelfth grade math test scores in Altonji, Elder, and Taber (2005) and Wooldridge (2015) is consistent with this scenario. In this study, the proportion of treated units is again small, and hence OLS and 2SLS can be used

[16] The implementation of these theoretical results in Stata is discussed in Appendix E.

to approximate ATT but not ATE. In this scenario, if the researcher only reports OLS or 2SLS estimates, she might give a reasonable answer to one of her questions but the other question will remain unanswered.

Finally, in the fourth scenario, my diagnostics indicate that—in the presence of treatment effect heterogeneity—OLS or 2SLS estimates might be quite different from those of our target parameter or parameters. Because treatment effects are indeed heterogeneous, our conclusions from the study are affected by the implicit OLS/2SLS weights; in other words, "bias from heterogeneity" is present. To provide an empirical illustration of this scenario, I replicate the analysis of the effects of cash transfers to poor families on longevity in Aizer *et al.* (2016). In this application, there are relatively many treated units and hence OLS places a disproportionately large weight on the effect on the controls. Also, this effect seems to be much larger than the effect on the treated, so OLS overestimates both ATE and ATT.

Scenario 1: OLS/2SLS Recovers the Parameter of Interest Despite the Presence of Heterogeneity

In their influential book, Angrist and Pischke (2009) reanalyze the NSW–CPS data, previously studied by LaLonde (1986), Dehejia and Wahba (1999), Smith and Todd (2005), and many others.[17] Angrist and Pischke (2009) conclude that OLS estimation of the model in (1) is a sensible strategy even in the presence of treatment effect heterogeneity, as their estimates of the effects of NSW program on earnings are reasonably similar to the known experimental estimate of 1,794. My analysis suggests that this conclusion is driven by the small proportion of treated units in the NSW–CPS sample.

In this context, it is clear that the researcher is only interested in the effect on the treated. The reason is simple: respondents from the CPS, who constitute the nonexperimental comparison group, are very different from individuals eligible for the program. What follows, it would not be reasonable to expect that any estimates of the effect on the controls could replicate the experimental benchmark, as this benchmark corresponds to the effect of NSW program on a different group of individuals.

If the researcher is primarily interested in the effect on the treated, then Corollary 4 suggests using $\hat{w}_0$, the estimated OLS weight on the effect on the controls, as a simple diagnostic. Then, if $\hat{w}_0$ is close to zero, OLS estimates will be similar to the implicit estimates of the effect on the treated. Table 2 reproduces the OLS estimates from Angrist

[17] More precisely, Angrist and Pischke (2009) analyze the subsample of the experimental treated units constructed by Dehejia and Wahba (1999), combined with "CPS-1" or "CPS-3," *i.e.* two of the nonexperimental comparison groups constructed by LaLonde (1986). In this replication, I focus on "CPS-1."

Table 2: OLS Estimates of the Effects of NSW Program

	(1)	(2)	(3)	(4)
	Estimates			
NSW program	–3,437***	–78	623	794
	(612)	(596)	(610)	(619)
	Diagnostics			
$\hat{w}_0$	0.019	0.001	0.017	0.017
$\hat{w}_0^* = \hat{\rho}$	0.011	0.011	0.011	0.011
$\hat{\delta}$	–0.970	–0.987	–0.971	–0.971
$\hat{\delta}^* = 2\hat{\rho} - 1$	–0.977	–0.977	–0.977	–0.977
Demographic controls	✓		✓	✓
"Earnings in 1974"				✓
Earnings in 1975		✓	✓	✓
Observations	16,177	16,177	16,177	16,177

Notes: These estimates correspond to column 2 in Table 3.3.3 in Angrist and Pischke (2009, p. 89). See also LaLonde (1986), Dehejia and Wahba (1999), and Smith and Todd (2005) for more details on these data. The dependent variable is earnings in 1978. Demographic controls include age, age squared, years of schooling, and indicators for married, high school dropout, black, and Hispanic. For treated individuals, "Earnings in 1974" correspond to real earnings in months 13–24 prior to randomization, which overlaps with calendar year 1974 for a number of individuals. See Smith and Todd (2005) for further discussion. Formulas for w_0 and δ are given in Theorem 1 and Corollary 2, respectively. Huber–White standard errors are in parentheses.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

and Pischke (2009) and reports the values of my diagnostics, with different columns corresponding to different sets of control variables. It turns out that $\hat{w}_0$ is between 0.1% and 1.9% for all specifications; similarly, the "rule of thumb" value of this diagnostic, $\hat{w}_0^*$, is, as always, equal to the proportion of treated units (only 1.1% in this sample). These results are very simple to interpret. Namely, we estimate that the difference between the OLS estimand and the effect on the treated is smaller than 2% of the difference between the effect on the controls and the effect on the treated. In this case, it might indeed be sensible to report only the OLS estimates of the effect of NSW program.

Table 3 provides an application of Theorem 1 to the OLS estimates in Table 2. As a result, all the baseline coefficients from Angrist and Pischke (2009) are now decomposed into two components, ATT and ATC. The difference between these estimates is always quite large. In column 4, while the estimate of the effect on the treated is 928, the effect on the controls is estimated to be –6,840. In other words, the OLS estimate of 794, reported in Angrist and Pischke (2009), is actually a weighted average of these two estimates. The

Table 3: NSW Program and Treatment Effect Heterogeneity

	(1)	(2)	(3)	(4)
NSW program	–3,437***	–78	623	794
	(612)	(596)	(610)	(619)
Decomposition (Theorem 1)				
$a.$ ATT	–3,373***	–69	754	928
	(636)	(598)	(672)	(673)
$b.$ $\hat{w}_1$	0.981	0.999	0.983	0.983
$c.$ ATC	–6,753***	–6,289**	–6,841***	–6,840***
	(1,201)	(2,561)	(1,271)	(1,298)
$d.$ $\hat{w}_0$	0.019	0.001	0.017	0.017
OLS $= a \cdot b + c \cdot d$	–3,437***	–78	623	794
	(612)	(596)	(610)	(619)
$e.$ $\hat{\mathrm{P}}\left(d=1\right)$	0.011	0.011	0.011	0.011
$f.$ $\hat{\mathrm{P}}\left(d=0\right)$	0.989	0.989	0.989	0.989
ATE $= a \cdot e + c \cdot f$	–6,714***	–6,218**	–6,754***	–6,751***
	(1,189)	(2,534)	(1,258)	(1,285)
Demographic controls	✓		✓	✓
"Earnings in 1974"				✓
Earnings in 1975		✓	✓	✓
Observations	16,177	16,177	16,177	16,177

Notes: See also LaLonde (1986), Dehejia and Wahba (1999), and Smith and Todd (2005) for more details on these data. The dependent variable is earnings in 1978. Demographic controls include age, age squared, years of schooling, and indicators for married, high school dropout, black, and Hispanic. For treated individuals, "Earnings in 1974" correspond to real earnings in months 13–24 prior to randomization, which overlaps with calendar year 1974 for a number of individuals. See Smith and Todd (2005) for further discussion. Estimates of ATE, ATT, and ATC are sample analogues of τ_{APE}, $\tau_{APE|d=1}$, and $\tau_{APE|d=0}$, respectively. Formulas for w_0 and w_1 are given in Theorem 1. Huber–White standard errors are in parentheses. Standard errors for ATE, ATT, and ATC ignore that the propensity score is estimated.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

fact that this estimate is close to 928, and not to –6,840, is a consequence of the small proportion of treated units in this sample, 1.1%. The weight on 928 is 98.3% and the weight on –6,840 is only 1.7%, as already reported in Table 2. If the proportion of treated units was larger, the weight on ATT would be smaller and the "performance" of OLS in replicating the experimental benchmark would deteriorate.

As a robustness check, I report a number of alternative estimates of the effects of NSW program in Appendix F. I consider nearest-neighbor matching on the LPM and logit propensity scores as well as the "Oaxaca–Blinder" estimator of average treatment effects. In each case, I separately estimate ATE, ATT, and ATC. These additional estimates are consistent with the claim that the general pattern of results in Table 3 is driven by the OLS weights. The estimates of ATE and ATC are always negative and large in magnitude; the estimates of ATT are much closer to the experimental benchmark.

Scenario 2: OLS/2SLS Recovers the Parameter(s) of Interest Due to Lack of Heterogeneity

A recent paper by Voigtländer and Voth (2012) examines the long-term persistence of cultural traits, focusing on anti-Semitic sentiment and violence in Germany. In particular, the authors analyze the data on over 300 towns where Jewish communities are documented for both the medieval and interwar periods; they demonstrate that medieval pogroms, which occurred during the Black Death (1348–50), are associated with higher levels of anti-Semitism in the 1920s and 1930s, as measured by pogroms, NSDAP votes in 1928, DVFP votes in 1924, attacks on synagogues during the *Reichskristallnacht* in 1938, and other expressions of anti-Semitic hatred.

In this study, the "treatment" variable equals one for towns where a pogrom occurred in the years 1348–50 and zero otherwise. It seems plausible that the researcher might be interested either in the average effect of medieval pogroms, ATE, or in the average effect of "treatment" on towns where a pogrom occurred, ATT. One might also argue that this latter parameter is more informative.

Table 4 reproduces the baseline estimates from Voigtländer and Voth (2012) and reports the values of my diagnostics. These results suggest caution in interpreting the OLS estimates. The proportion of towns where a Black Death pogrom occurred is very high and equal to 72.3–75.9%. What follows, we expect OLS to place a disproportionately large weight—of the same order of magnitude—on the effect on the controls. In fact, we estimate that the difference between the OLS estimand and the effect on the treated is between 71.5% and 76.2% of the difference between the effect on the controls and the effect

Table 4: OLS Estimates of the Effects of Medieval Pogroms

	1920s pogroms	NSDAP 1928	DVFP 1924	Synagogue attacks
	Estimates			
Medieval pogrom	0.0607***	0.0142**	0.0147	0.1239**
	(0.0226)	(0.0057)	(0.0110)	(0.0522)
	Diagnostics			
$\hat{w}_0$	0.734	0.715	0.733	0.762
$\hat{w}_0^* = \hat{\rho}$	0.725	0.723	0.723	0.759
$\hat{\delta}$	0.459	0.438	0.456	0.521
$\hat{\delta}^* = 2\hat{\rho} - 1$	0.450	0.446	0.446	0.518
Log of population in 1924			✓	
Log of population in 1925	✓			
Log of population in 1928		✓		
Log of population in 1933				✓
% Jewish in 1925	✓	✓	✓	
% Jewish in 1933				✓
% Protestant in 1925	✓	✓	✓	✓
Observations	320	325	325	278

Notes: These estimates correspond to columns 1, 2, 3, and 6 in Table VI in Voigtländer and Voth (2012, p. 1365). The dependent variables are an indicator for pogroms during the 1920s, the vote share of the NSDAP in the May 1928 election, the vote share of the Deutsch-Völkische Freiheitspartei in the May 1924 election, and an indicator for whether a synagogue was destroyed or damaged in the 1938 *Reichskristallnacht*. Formulas for w_0 and δ are given in Theorem 1 and Corollary 2, respectively. Cluster-robust standard errors are in parentheses. *Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

on the treated. The difference between OLS and ATE is estimated to be between 43.8% and 52.1% of this measure of heterogeneity. In consequence, if the effects of medieval pogroms on modern anti-Semitism were really heterogeneous, the OLS estimates would be very biased.[18]

Further details on the heterogeneity in these effects are presented in Table 5, which applies Theorem 1 to the baseline estimates in Table 4. What follows, each OLS estimate

[18] This suggestion is mostly illustrative, as Voigtländer and Voth (2012) also report matching estimates of the effects of medieval pogroms, which are similar to the OLS estimates. Moreover, the authors consider two other measures of modern anti-Semitism, the number of deportations of Jews after 1933 and the number of anti-Semitic letters to *Der Stürmer* per 10,000 inhabitants. I ignore these measures for two reasons. First, in these two cases, Voigtländer and Voth (2012) use the Poisson regression model instead of OLS estimation of the model in (1). Second, the effects of medieval pogroms are more heterogeneous in the case of these two measures, and hence they are inconsistent with my description of "scenario 2" that applied researchers are likely to face. Nevertheless, my results for these measures are also mostly consistent with Voigtländer and Voth (2012).

Table 5: Medieval Pogroms and Treatment Effect Heterogeneity

	1920s pogroms	NSDAP 1928	DVFP 1924	Synagogue attacks
Medieval pogrom	0.0607***	0.0142**	0.0147	0.1239**
	(0.0226)	(0.0057)	(0.0110)	(0.0522)
Decomposition (Theorem 1)				
a. ATT	0.0563*	0.0115**	0.0122	0.1150**
	(0.0292)	(0.0052)	(0.0110)	(0.0470)
b. $\hat{w}_1$	0.266	0.285	0.267	0.238
c. ATC	0.0623***	0.0153***	0.0156	0.1267**
	(0.0207)	(0.0057)	(0.0112)	(0.0548)
d. $\hat{w}_0$	0.734	0.715	0.733	0.762
OLS $= a \cdot b + c \cdot d$	0.0607***	0.0142**	0.0147	0.1239**
	(0.0226)	(0.0057)	(0.0110)	(0.0522)
e. $\hat{P}(d = 1)$	0.725	0.723	0.723	0.759
f. $\hat{P}(d = 0)$	0.275	0.277	0.277	0.241
ATE $= a \cdot e + c \cdot f$	0.0579**	0.0125**	0.0132	0.1178**
	(0.0264)	(0.0053)	(0.0109)	(0.0484)
Log of population in 1924			✓	
Log of population in 1925	✓			
Log of population in 1928		✓		
Log of population in 1933				✓
% Jewish in 1925	✓	✓	✓	
% Jewish in 1933				✓
% Protestant in 1925	✓	✓	✓	✓
Observations	320	325	325	278

Notes: See also Voigtländer and Voth (2012) for more details on these data. The dependent variables are an indicator for pogroms during the 1920s, the vote share of the NSDAP in the May 1928 election, the vote share of the Deutsch-Völkische Freiheitspartei in the May 1924 election, and an indicator for whether a synagogue was destroyed or damaged in the 1938 *Reichskristallnacht*. Estimates of ATE, ATT, and ATC are sample analogues of τ_{APE}, $\tau_{APE|d=1}$, and $\tau_{APE|d=0}$, respectively. Formulas for w_0 and w_1 are given in Theorem 1. Cluster-robust standard errors are in parentheses. Standard errors for ATE, ATT, and ATC ignore that the propensity score is estimated.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

from Voigtländer and Voth (2012) is now represented as a weighted average of two other estimates, for towns where a medieval pogrom occurred (ATT) and for towns where a medieval pogrom did not occur (ATC). There is no evidence of meaningful treatment effect heterogeneity. On the one hand, the estimates of the effect on the controls are consistently larger than those of the effect on the treated. Consequently, since $\hat{\delta} > 0$ (which, in turn, is due to the fact that $\hat{\rho} > 50\%$), OLS also seems to overestimate the average effect of medieval pogroms (ATE). On the other hand, these differences are very small and little is lost by reporting only the OLS estimates.

Additionally, I provide a number of alternative estimates of the effects of medieval pogroms in Appendix F. These results confirm that the effects on the treated and controls are similar. What follows, my replication of Voigtländer and Voth (2012) illustrates a simple fact: when treatment effects are homogeneous, the OLS weights do not matter in practice; even if their distribution between ATT and ATC is potentially quite harmful, a "weird" convex combination of two similar objects remains similar to both of them.

Scenario 3: OLS/2SLS Recovers One of Two Parameters of Interest

In the next empirical illustration I replicate some of the estimates of the effects of Catholic schooling on math test scores from Wooldridge (2015) who, in turn, revisits an earlier study by Altonji *et al.* (2005). In this case, we use distance from the student's home to the nearest Catholic high school as an instrument for Catholic schooling. Hence, in this application, I represent the OLS and 2SLS estimates from Wooldridge (2015) as weighted averages of estimates of ATT and ATC. It is important to note that Wooldridge (2015) also uses a number of control function approaches to estimate various parameters of interest, including ATT and ATC, so this replication is mostly illustrative.

In the context of the effectiveness of various institutions or policies, it is often interesting to make inferences about both ATE and ATT. In a study of the effects of Catholic schooling, the difference between these parameters is informative about whether students select to Catholic schools on the basis of gains from such schooling. In general, we might expect that ATT is larger than ATE (or ATC). If the effect on the treated is indeed larger than the effect on the controls, then an implication of Corollary 4 is that OLS (and 2SLS) will be negatively biased for ATT. The direction of bias in estimating ATE is data dependent but can be inferred from the sign of δ (Corollary 2) or δ_{ts} (footnote 14).

Table 6 reproduces an OLS and a 2SLS estimate from Wooldridge (2015) in columns 1 and 3; it also reports estimates which additionally use a number of demographic control variables. Finally, Table 6 reports the values of my diagnostics. Since $\hat{\delta}$ and $\hat{\delta}_{ts}$ are always

Table 6: OLS and 2SLS Estimates of the Effects of Catholic Schooling

	(1)	(2)	(3)	(4)
	OLS		2SLS	
	Estimates			
Catholic high school	1.49***	1.61***	2.36*	3.36***
	(0.39)	(0.38)	(1.25)	(1.22)
	Diagnostics			
$\hat{w}_0$ / $\hat{w}_{0,ts}$	0.055	0.058	0.090	0.086
$\hat{w}_0^*$ / $\hat{w}_{0,ts}^*$ $(= \hat{\rho})$	0.061	0.061	0.061	0.061
$\hat{\delta}$ / $\hat{\delta}_{ts}$	–0.884	–0.882	–0.849	–0.854
$\hat{\delta}^*$ / $\hat{\delta}_{ts}^*$ $(= 2\hat{\rho} - 1)$	–0.879	–0.878	–0.879	–0.878
Baseline controls	✓	✓	✓	✓
Demographic controls		✓		✓
Observations	7,444	7,306	7,444	7,306

Notes: The estimates in columns 1 and 3 correspond to columns 1 and 2 in Table 2 in Wooldridge (2015, p. 426). See also Altonji *et al.* (2005) for more details on these data. The dependent variable is a standardized twelfth grade math test score. Baseline controls include mother's education, father's education, and log of family income. Demographic controls include indicators for female, Asian, black, Hispanic, a married parent, and a single-mother household. In columns 3 and 4, instrumental variables for Catholic schooling include the following indicators for distance from the nearest Catholic high school: less than 1 mile, between 1 and 3 miles, between 3 and 6 miles, and between 6 and 10 miles. Formulas for w_0, $w_{0,ts}$, δ, and δ_{ts} are given in Theorem 1, Theorem 3, Corollary 2, and footnote 14, respectively. Huber–White standard errors are in parentheses.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

negative, we expect OLS and 2SLS to be positively biased for ATE, as long as ATT is also larger than ATC. In fact, this bias can be quite substantial, as it is estimated to correspond to 84.9–88.4% of the difference between ATT and ATC. On the other hand, the bias in estimating ATT will be small; it is expected to be equal to 5.5–9.0% of this difference. Neither of these claims should be surprising, however, since the proportion of treated units, 6.1%, is also very small. What follows, if the effects of Catholic schooling are heterogeneous and the researcher reports only the OLS and 2SLS estimates, she will likely gain (approximate) knowledge of the average treatment effect on the treated but not of the average treatment effect. This latter parameter is implicitly "swept under the rug."

The results in Table 7, which applies Theorems 1 and 3 to the estimates in Table 6, are consistent with this picture. The use of instrumental variables in columns 3 and 4 reveals a substantial amount of heterogeneity. Both estimates of the effect on the treated are positive and statistically significant; the estimates of the effect on the controls are negative but not statistically different from zero. In consequence, the 2SLS estimates are also positive

Table 7: Catholic Schooling and Treatment Effect Heterogeneity

	(1)	(2)	(3)	(4)
	OLS		2SLS	
Catholic high school	1.49***	1.61***	2.36*	3.36***
	(0.39)	(0.38)	(1.25)	(1.22)
Decomposition (Theorems 1 and 3)				
a. ATT	1.47***	1.58***	2.89**	3.93***
	(0.40)	(0.41)	(1.44)	(1.45)
b. $\hat{w}_1$ / $\hat{w}_{1,ts}$	0.945	0.942	0.910	0.914
c. ATC	1.92***	2.13***	–2.91	–2.77
	(0.45)	(0.46)	(3.88)	(4.04)
d. $\hat{w}_0$ / $\hat{w}_{0,ts}$	0.055	0.058	0.090	0.086
OLS / 2SLS $= a \cdot b + c \cdot d$	1.49***	1.61***	2.36*	3.36***
	(0.39)	(0.38)	(1.25)	(1.22)
e. $\hat{\mathrm{P}}\left(d=1\right)$	0.061	0.061	0.061	0.061
f. $\hat{\mathrm{P}}\left(d=0\right)$	0.939	0.939	0.939	0.939
ATE $= a \cdot e + c \cdot f$	1.89***	2.09***	–2.56	–2.37
	(0.44)	(0.45)	(3.65)	(3.80)
Baseline controls	✓	✓	✓	✓
Demographic controls		✓		✓
Observations	7,444	7,306	7,444	7,306

Notes: See also Altonji *et al.* (2005) for more details on these data. The dependent variable is a standardized twelfth grade math test score. Baseline controls include mother's education, father's education, and log of family income. Demographic controls include indicators for female, Asian, black, Hispanic, a married parent, and a single-mother household. In columns 3 and 4, instrumental variables for Catholic schooling include the following indicators for distance from the nearest Catholic high school: less than 1 mile, between 1 and 3 miles, between 3 and 6 miles, and between 6 and 10 miles. Estimates of ATE, ATT, and ATC are sample analogues of τ_{APE} or $\tau_{APE,ts}$, $\tau_{APE|d=1}$ or $\tau_{APE,ts|d=1}$, and $\tau_{APE|d=0}$ or $\tau_{APE,ts|d=0}$, respectively. Formulas for w_0 and w_1 ($w_{0,ts}$ and $w_{1,ts}$) are given in Theorem 1 (Theorem 3). Huber–White standard errors are in parentheses. Standard errors for ATE, ATT, and ATC ignore that the propensity score is estimated.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

and quite large. While they provide a reasonable approximation to the implicit estimates of ATT, they also ignore the fact that the estimates of ATE are very different. There is no evidence that a randomly drawn individual would benefit from Catholic schooling.

As a robustness check, I report a number of alternative estimates of the effects of Catholic schooling on math test scores in Appendix F. In the case of instrumental variables estimation, as in Wooldridge (2015), I estimate the Gaussian switching regime model using the Heckman's two-step procedure. It turns out that these estimates are similar to the implicit 2SLS estimates of ATE, ATT, and ATC in Table 7, although the estimates of ATC (and hence ATE) in Appendix F are also closer to zero.

Scenario 4: OLS/2SLS Does Not Recover the Parameter(s) of Interest

In my final empirical application, I follow a recent paper by Aizer *et al.* (2016) and study the effects of cash transfers to poor families on longevity of the children of their beneficiaries. In particular, Aizer *et al.* (2016) analyze the administrative records of applicants to the Mothers' Pension (MP) program, which was the first welfare program sponsored by the U.S. government. It supported poor mothers with dependent children and was launched on a state-by-state basis between 1911 and 1931.

In this study, the control group consists only of children of mothers who applied to the program, were initially deemed eligible, but were ultimately rejected. This strategy is used to ensure that treated and control individuals are broadly comparable; nevertheless, at the time of application, rejected mothers were somewhat better-off than accepted mothers. As before, it seems plausible that the researcher might be interested either in the average effect of cash transfers, ATE, or in the average effect of cash transfers on accepted applicants, ATT. The former parameter would be particularly interesting in the context of a hypothetical expansion of the program. On the other hand, the latter parameter is a simple measure of the effects of this program on its actual participants. In fact, one might argue that in this context the average treatment effect on the treated is intuitively more appealing as a target parameter.

Table 8 reproduces the baseline estimates from Aizer *et al.* (2016) and reports the values of my diagnostics. While the OLS estimates are positive and statistically significant, my diagnostics indicate that these results should be approached with caution. Namely, treated units constitute the majority (or 87.5%) of the sample. What follows, we expect OLS to be potentially very biased for both ATE and ATT (see Corollaries 2 to 4). Indeed, my estimates of δ suggest that the difference between the OLS estimand and the average treatment effect is equal to 65.9–74.5% of the difference between ATC and ATT. The esti-

Table 8: OLS Estimates of the Effects of Cash Transfers

	(1)	(2)	(3)	(4)
		Estimates		
MP program	0.0157**	0.0158**	0.0182**	0.0167**
	(0.0065)	(0.0065)	(0.0069)	(0.0072)
		Diagnostics		
$\hat{w}_0$	0.861	0.870	0.784	0.784
$\hat{w}_0^* = \hat{\rho}$	0.875	0.875	0.875	0.875
$\hat{\delta}$	0.736	0.745	0.659	0.659
$\hat{\delta}^* = 2\hat{\rho} - 1$	0.750	0.750	0.750	0.750
State fixed effects	✓			
County fixed effects			✓	✓
Cohort fixed effects	✓	✓	✓	✓
State characteristics		✓	✓	✓
County characteristics		✓		
Individual characteristics		✓	✓	✓
Observations	7,860	7,859	7,859	7,857

Notes: These estimates correspond to columns 1 to 4 in panel A of Table 4 in Aizer *et al.* (2016, p. 952). The dependent variable is log age at death, as reported in the MP records (columns 1 to 3) or on the death certificate (column 4). State characteristics include manufacturing wages, age of school entry, minimum age for work permit, an indicator for a continuation school requirement, state laws concerning MP transfers (work requirement, reapplication requirement, and maximum amounts for first and second child), and log expenditures on education, charity, and social programs. County characteristics include average value of farm land, mean and SD of socio-economic index, poverty rate, female lfp rate, and shares of urban population, widowed women, children living with single mothers, and children working. Individual characteristics include child age at application, age of oldest and youngest child in family, number of letters in name, and indicators for the number of siblings, the marital status of the mother, and whether date of birth is incomplete. Formulas for w_0 and δ are given in Theorem 1 and Corollary 2, respectively. Cluster-robust standard errors are in parentheses.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

mates of w_0 suggest that the difference between OLS and ATT corresponds to 78.4–87.0% of this measure of heterogeneity, which is—and should typically be—similar to the proportion of treated units. It turns out that in the presence of treatment effect heterogeneity the OLS estimates of the effects of cash transfers on longevity might be substantially biased for both of our parameters of interest.

The results in Table 9 suggest that this might indeed be the case. In Table 9, following Theorem 1, each OLS estimate from Table 8 is represented as a weighted average of two other estimates, on accepted applicants (ATT) and on rejected applicants (ATC). The estimates of the effect on the controls are consistently larger than those of the effect on the treated. Thus, OLS overestimates both ATE (since $\hat{\delta} > 0$) and ATT. While my estimates of

Table 9: Cash Transfers and Treatment Effect Heterogeneity

	(1)	(2)	(3)	(4)
MP program	0.0157**	0.0158**	0.0182**	0.0167**
	(0.0065)	(0.0065)	(0.0069)	(0.0072)
Decomposition (Theorem 1)				
a. ATT	0.0129*	0.0149*	0.0097	0.0089
	(0.0076)	(0.0081)	(0.0093)	(0.0093)
b. $\hat{w}_1$	0.139	0.130	0.216	0.216
c. ATC	0.0162**	0.0160**	0.0206**	0.0188**
	(0.0079)	(0.0077)	(0.0084)	(0.0085)
d. $\hat{w}_0$	0.861	0.870	0.784	0.784
OLS $= a \cdot b + c \cdot d$	0.0157**	0.0158**	0.0182**	0.0167**
	(0.0065)	(0.0065)	(0.0069)	(0.0072)
e. $\hat{\mathrm{P}}\left(d = 1\right)$	0.875	0.875	0.875	0.875
f. $\hat{\mathrm{P}}\left(d = 0\right)$	0.125	0.125	0.125	0.125
ATE $= a \cdot e + c \cdot f$	0.0133*	0.0150*	0.0110	0.0102
	(0.0076)	(0.0078)	(0.0088)	(0.0088)
State fixed effects	✓			
County fixed effects			✓	✓
Cohort fixed effects	✓	✓	✓	✓
State characteristics		✓	✓	✓
County characteristics		✓		
Individual characteristics		✓	✓	✓
Observations	7,860	7,859	7,859	7,857

Notes: See also Aizer *et al.* (2016) for more details on these data. The dependent variable is log age at death, as reported in the MP records (columns 1 to 3) or on the death certificate (column 4). State characteristics include manufacturing wages, age of school entry, minimum age for work permit, an indicator for a continuation school requirement, state laws concerning MP transfers (work requirement, reapplication requirement, and maximum amounts for first and second child), and log expenditures on education, charity, and social programs. County characteristics include average value of farm land, mean and SD of socio-economic index, poverty rate, female lfp rate, and shares of urban population, widowed women, children living with single mothers, and children working. Individual characteristics include child age at application, age of oldest and youngest child in family, number of letters in name, and indicators for the number of siblings, the marital status of the mother, and whether date of birth is incomplete. Estimates of ATE, ATT, and ATC are sample analogues of τ_{APE}, $\tau_{APE|d=1}$, and $\tau_{APE|d=0}$, respectively. Formulas for w_0 and w_1 are given in Theorem 1. Cluster-robust standard errors are in parentheses. Standard errors for ATE, ATT, and ATC ignore that the propensity score is estimated.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

these parameters remain statistically significant in columns 1 and 2, this is no longer the case in columns 3 and 4, following the inclusion of county fixed effects. Perhaps more importantly, these estimates of ATT are half smaller than the corresponding OLS estimates. Clearly, this difference is economically quite large.

To assess the robustness of these findings, I report a number of alternative estimates of the effects of cash transfers in Appendix F. These additional results seem to reinforce my conclusion. Only one in twelve estimates of the effect on the treated is statistically different from zero, and four of the insignificant estimates are negative. While it is entirely possible—and perhaps quite likely—that participation in the MP program increased longevity of the children of its beneficiaries, the OLS estimates in Aizer *et al.* (2016) are almost certainly too large. Interestingly, this bias is driven by the implicit OLS weights on the effects on the treated and controls; these weights turn out to be particularly counterproductive in this application.

4 Extensions

This section discusses several further extensions of my main theoretical results. In particular, I discuss the implications of my baseline result for regression adjustments to experimental data, fixed effects, difference-in-differences estimation, and the interpretation of the IV estimand as the local average treatment effect.

Regression Adjustments to Experimental Data

It is important to note that Theorem 1 does not apply to two simplest settings in the literature on regression adjustments to experimental data: Bernoulli trials and completely randomized experiments. In these cases, discussed in detail by Imbens and Rubin (2015), each unit has the same probability of treatment assignment. Consequently, $\mathrm{V}\left[p\left(X\right)\right] = \mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right] = 0$, and hence Assumption 2 is not satisfied and Theorem 1 does not apply. Even if it were to apply, however, it would not suggest any problems with regression adjustments. The reason is simple: in these two settings, $\tau_{ATT} = \tau_{ATC}$, and so every convex combination of these two objects is the same. In fact, Imbens and Rubin (2015) confirm that—when the data come from a Bernoulli trial or a completely randomized experiment—OLS estimation of the model in (1) is consistent for τ_{ATE}.

In other experimental settings, however, problems can arise. In particular, Imbens and Rubin (2015, Ch. 9.6) show that—in a stratified randomized experiment—regression adjustments with strata fixed effects will be inconsistent for τ_{ATE}, which is a straightforward application of the result in Angrist (1998). In this case, Assumption 2 is satisfied and Theorem 1 also applies. On the other hand, Bugni, Canay, and Shaikh (2017) demonstrate that in an important special case—when treatment assignment probabilities are equal in each stratum—this inconsistency disappears. As before, in this special case, $\mathrm{V}\left[p\left(X\right)\right] = \mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right] = 0$. Thus, again, Assumption 2 is not satisfied and Theorem 1 does not apply. Also, in this special case, unlike in the general case studied by Imbens and Rubin (2015, Ch. 9.6), $\tau_{ATT} = \tau_{ATC}$, so there can be little reason to expect inconsistency anyway.

Fixed Effects

On the other hand, Theorem 1 does, in fact, apply to fixed effects (FE) estimation of the basic unobserved effects model, which is similar to (1), but it also includes an unobserved component ("fixed effect") for each cross section observation. To see this, note that our estimand of interest is the same whether we consider FE estimation of the unobserved effects model or OLS estimation of a model where each unobserved component is treated as an unknown parameter to be estimated (*i.e.* the least squares dummy variable model). In other words, we are interested in the interpretation of τ_{fe} in

$$\mathrm{L}\left(y \mid d, X_{fe}, c_1, \ldots, c_N\right) = \tau_{fe} d + X_{fe}\beta_{fe} + \sum_{i=1}^{N} \alpha_{fe,i} c_i, \tag{27}$$

where $c_1, \ldots, c_N$ are indicators for cross section observations; the remaining covariates are denoted by X_{fe}, and hence $X = \left(X_{fe}, c_1, \ldots, c_N\right)$.

Clearly, Theorem 1 can be applied directly to equation (27). What follows, τ_{fe} is also identical to the outcome of a particular three-step procedure. In the first step, we obtain $p\left(X\right)$ from a linear projection of d on X_{fe} and the indicator variables, $c_1, \ldots, c_N$. In the second step, we obtain $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$, as in (11), from two linear projections of y on $p\left(X\right)$, separately for $d = 1$ and $d = 0$. In the third step, we calculate a weighted average of $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$, where the weights, again, follow from Theorem 1.

The fact that τ_{fe} is not, in general, equal to τ_{ATE} is well established. However, this particular representation of the FE estimand is new. It differs from and complements the earlier literature, such as Wooldridge (2005) and Gibbons *et al.* (2018).

Difference-In-Differences

It is also useful to discuss the implications of the results in this paper for difference-in-differences (DD) estimation. In the simplest case, we consider two groups, treated and controls, and two time periods. No units receive treatment in the first time period. In the second time period, treated units are exposed to treatment, while control units are not. Consequently, we are interested in the interpretation of τ_{dd} in

$$\mathrm{L}\left(y \mid 1, d, g, t\right) = \alpha_{dd} + \tau_{dd} d + \chi_{dd} g + \psi_{dd} t, \tag{28}$$

where g is an indicator for those in the treatment group, t is an indicator for the second time period, and $d = g \cdot t$ is our interaction term of interest; also, $X = (g, t)$. Thus, if we apply Theorem 1 to τ_{dd}, our "propensity score" has the form

$$p\left(X\right) = \mathrm{L}\left(d \mid 1, g, t\right) = \alpha_{p,dd} + \chi_{p,dd} g + \psi_{p,dd} t. \tag{29}$$

It follows immediately that in this case $\mathrm{V}\left[p\left(X\right) \mid d = 1\right] = 0$ but $\mathrm{V}\left[p\left(X\right) \mid d = 0\right] > 0$. Whenever $d = 1$, $p\left(X\right) = \alpha_{p,dd} + \chi_{p,dd} + \psi_{p,dd}$, and hence for these units there is no variation in the "propensity score." What follows, Assumption 2 is not satisfied and Theorem 1 does not apply. It is possible, however, to relax Assumption 2.

Assumption 5 $\mathrm{V}\left[p\left(X\right) \mid d = 1\right]$ *is zero.* $\mathrm{V}\left[p\left(X\right) \mid d = 0\right]$ *is nonzero.*

Assumption 5 allows one of the conditional variances to be zero. Of course, the labeling of groups with $d = 1$ and $d = 0$ as "treated" and "controls" is arbitrary, so the zero conditional variance is allowed for the treated units or for the control units, but not for both of these groups. Under Assumptions 1 and 5, a new result can be derived.

Theorem 4 (Difference-In-Differences Estimation) *Suppose that Assumptions 1 and 5 are satisfied. Then,*

$$\tau = \tau_{APE|d=1}.$$

The proof of Theorem 4 follows directly from the proof of Theorem 1. According to Theorem 4, if the "propensity score" does not vary in one of the groups, the OLS estimand will be equal to the effect on this group. As an example, this result is applicable to DD estimation. In this context, it is well known that a model without additional covariates, as in (28), permits identification of the effect on the treated (see, *e.g.*, Heckman, Ichimura, Smith, and Todd, 1998; Abadie, 2005).[19] Theorem 4 reaches the same conclusion.

If, however, X contains additional variables, say $X = (g, t, X_{dd})$, then Theorem 4 will no longer be applicable, as both conditional variances will be nonzero. Instead, we will be able to use Theorem 1. What follows, with additional covariates and without interactions between these covariates and d, the DD estimand is equal to a convex combination of $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$, and not simply to $\tau_{APE|d=1}$. Empirically, however, w_1 might often be close to one. First, in many applications of the difference-in-differences approach, there are few treated units and many control units; consequently, w_1 is also relatively large. Second, in most applications, we might expect that $\mathrm{V}\left[p\left(X\right) \mid d=0\right] \gg \mathrm{V}\left[p\left(X\right) \mid d=1\right]$, as g and t should have more explanatory power for $d = g \cdot t$ than X_{dd}; again, this would increase the OLS weight on the effect on the treated, w_1.

Local Average Treatment Effects

A common interpretation of the instrumental variables estimand as the local average treatment effect, *i.e.* the average effect on those individuals whose treatment status is affected by the instrumental variable (compliers), is motivated by the results in Imbens and Angrist (1994). It is often overlooked, however, that these results apply only to models without covariates, X. On the other hand, Angrist and Imbens (1995) extend this interpretation to saturated models with covariates, and demonstrate that the estimand of interest is a weighted average of stratum-specific effects. In Theorem 3, I provide an alternative weighted average representation of the 2SLS estimand. Hence, it might be useful to consider the implications of the results in this paper for the LATE interpretation of IV.

This discussion is facilitated by focusing, as is often the case in the literature on local average treatment effects, on a model with a single binary instrumental variable, z. Hence, we are interested in the interpretation of τ_{iv} in

$$\mathrm{L}\left[y \mid 1, d\left(X, z\right), X\right] = \alpha_{iv} + \tau_{iv} \cdot d\left(X, z\right) + X\beta_{iv}, \tag{30}$$

[19] Recent papers by Borusyak and Jaravel (2017), Abraham and Sun (2018), Athey and Imbens (2018), de Chaisemartin and D'Haultfœuille (2018), Goodman-Bacon (2018), Hull (2018), and Strezhnev (2018) study the interpretation of the DD estimand in various extensions of the canonical 2x2 model. In general, these interpretations are substantially different than in the canonical case. However, in contrast to this paper, these authors concentrate most of their attention on models without additional covariates. See also Callaway and Sant'Anna (2018) for a recent study of identification and estimation of treatment effects using difference-in-differences with multiple time periods and variation in treatment timing, which also allows for conditioning on observed covariates.

where

$$d(X,z) = \mathrm{L}(d \mid 1, X, z) = \alpha_{rd} + X\beta_{rd} + \zeta_{rd}z. \tag{31}$$

Since z is now a scalar, we can also write τ_{iv}, the IV estimand, as $\tau_{iv} = \zeta_{ry}/\zeta_{rd}$, where ζ_{ry} is defined by

$$\mathrm{L}(y \mid 1, X, z) = \alpha_{ry} + X\beta_{ry} + \zeta_{ry}z. \tag{32}$$

Because z is also binary, we can provide a new interpretation of the IV estimand by applying Theorem 1 separately to ζ_{ry} and ζ_{rd}. As before, we need to begin with redefining several objects of interest. Indeed, let

$$\pi = \mathrm{P}(z = 1) \tag{33}$$

denote the unconditional probability that $z = 1$ and let

$$r(X) = \mathrm{L}(z \mid 1, X) = \alpha_{s,iv} + X\beta_{s,iv} \tag{34}$$

denote the "instrument propensity score." Also, we need to define two linear projections of y on $r(X)$, separately for $z = 1$ and $z = 0$, namely

$$\mathrm{L}[y \mid 1, r(X)] = \alpha_{1,y} + \gamma_{1,y} \cdot r(X) \quad \text{if} \quad z = 1 \tag{35}$$

and

$$\mathrm{L}[y \mid 1, r(X)] = \alpha_{0,y} + \gamma_{0,y} \cdot r(X) \quad \text{if} \quad z = 0, \tag{36}$$

as well as analogous linear projections of d on $r(X)$, namely

$$\mathrm{L}[d \mid 1, r(X)] = \alpha_{1,d} + \gamma_{1,d} \cdot r(X) \quad \text{if} \quad z = 1 \tag{37}$$

and

$$\mathrm{L}[d \mid 1, r(X)] = \alpha_{0,d} + \gamma_{0,d} \cdot r(X) \quad \text{if} \quad z = 0. \tag{38}$$

Finally, we can use the linear projections in (35) and (36) to define the average partial effect of z on y as

$$\tau_{APE,zy} = \left(\alpha_{1,y} - \alpha_{0,y}\right) + \left(\gamma_{1,y} - \gamma_{0,y}\right) \cdot \mathrm{E}[r(X)] \tag{39}$$

as well as the average partial effect of z on y in group k ($k = 0, 1$) as

$$\tau_{APE,zy|z=k} = \left(\alpha_{1,y} - \alpha_{0,y}\right) + \left(\gamma_{1,y} - \gamma_{0,y}\right) \cdot \mathrm{E}[r(X) \mid z = k]. \tag{40}$$

Similarly, using the linear projections in (37) and (38), we can define the average partial effect of z on d as

$$\tau_{APE,zd} = (\alpha_{1,d} - \alpha_{0,d}) + (\gamma_{1,d} - \gamma_{0,d}) \cdot \mathrm{E}\left[r\left(X\right)\right], \tag{41}$$

and the average partial effect of z on d in group k ($k = 0, 1$) as

$$\tau_{APE,zd|z=k} = (\alpha_{1,d} - \alpha_{0,d}) + (\gamma_{1,d} - \gamma_{0,d}) \cdot \mathrm{E}\left[r\left(X\right) \mid z = k\right]. \tag{42}$$

As before, the linear projections in (34) to (38) are definitional. We only require that these linear projections exist and are unique, and this is guaranteed by Assumptions 6 and 7.

Assumption 6 *(i)* $\mathrm{E}(y^2)$, $\mathrm{E}(d^2)$, *and* $\mathrm{E}(\|X\|^2)$ *are finite. (ii) The covariance matrices of* X *and* (X, z) *are nonsingular.*

Assumption 7 $\mathrm{V}\left[r\left(X\right) \mid z = 1\right]$ *and* $\mathrm{V}\left[r\left(X\right) \mid z = 0\right]$ *are nonzero.*

Similar to assumptions underlying Theorems 1 and 3, Assumptions 6 and 7 are generally innocuous, although Assumption 7 rules out the possibility that z is completely randomized and hence $r\left(X\right)$ is the same for all units. When z is completely randomized, the IV estimand is equal to the local average treatment effect, even if we also control for additional covariates. Finally, we require that z is relevant for treatment, *i.e.* that it is partially correlated with d.

Assumption 8 (Instrument Relevance) ζ_{rd} *is nonzero.*

Given that Assumptions 6, 7, and 8 are satisfied, the IV estimand is equal to the ratio of two weighted averages. The objects defined in (40) appear in the numerator and the objects defined in (42) appear in the denominator. This interpretation of the instrumental variables estimand follows from Theorem 5.

Theorem 5 (Interpretation of IV as a Ratio of Weighted Averages) *Suppose that Assumptions 6, 7, and 8 are satisfied. Then,*

$$\tau_{iv} = \frac{\pi \cdot \mathrm{V}\left[r\left(X\right) \mid z = 1\right] \cdot \tau_{APE,zy|z=0} + (1 - \pi) \cdot \mathrm{V}\left[r\left(X\right) \mid z = 0\right] \cdot \tau_{APE,zy|z=1}}{\pi \cdot \mathrm{V}\left[r\left(X\right) \mid z = 1\right] \cdot \tau_{APE,zd|z=0} + (1 - \pi) \cdot \mathrm{V}\left[r\left(X\right) \mid z = 0\right] \cdot \tau_{APE,zd|z=1}}.$$

The proof of Theorem 5 follows from the observation that $\tau_{iv} = \zeta_{ry} / \zeta_{rd}$. It is then sufficient to apply Theorem 1 separately to ζ_{ry} and ζ_{rd}, and simplify. In particular, $\zeta_{ry} = \frac{\pi \cdot \mathrm{V}[r(X)|z=1]}{\pi \cdot \mathrm{V}[r(X)|z=1] + (1-\pi) \cdot \mathrm{V}[r(X)|z=0]} \cdot \tau_{APE,zy|z=0} + \frac{(1-\pi) \cdot \mathrm{V}[r(X)|z=0]}{\pi \cdot \mathrm{V}[r(X)|z=1] + (1-\pi) \cdot \mathrm{V}[r(X)|z=0]} \cdot \tau_{APE,zy|z=1}$. Also, $\zeta_{rd} = \frac{\pi \cdot \mathrm{V}[r(X)|z=1]}{\pi \cdot \mathrm{V}[r(X)|z=1] + (1-\pi) \cdot \mathrm{V}[r(X)|z=0]} \cdot \tau_{APE,zd|z=0} + \frac{(1-\pi) \cdot \mathrm{V}[r(X)|z=0]}{\pi \cdot \mathrm{V}[r(X)|z=1] + (1-\pi) \cdot \mathrm{V}[r(X)|z=0]} \cdot \tau_{APE,zd|z=1}$.

After providing this interpretation of the IV estimand, it seems sensible to compare τ_{iv} to some benchmark, say

$$\tau_{APE,zy/zd} = \frac{\tau_{APE,zy}}{\tau_{APE,zd}}. \tag{43}$$

Note that $\tau_{APE,zy/zd}$ will be equal to the local average treatment effect if several conditions are satisfied.[20] Namely, we require that z is unconfounded conditional on covariates, X. Also, the population model for z needs to be linear in X and the population models for both potential outcomes and potential treatments need to be linear in $r(X)$. In general, according to Corollary 5, even if these assumptions are satisfied, we cannot expect τ_{iv} to recover the local average treatment effect.

Corollary 5 *Suppose that* $\mathrm{V}[r(X) \mid z=1] = \mathrm{V}[r(X) \mid z=0]$. *Then, Theorem 5 implies that*

$$\tau_{iv} = \tau_{APE,zy/zd} \qquad \textit{if and only if} \qquad \pi = 50\% \qquad \textit{or} \qquad \frac{\tau_{APE,zy|z=1}}{\tau_{APE,zd|z=1}} = \frac{\tau_{APE,zy|z=0}}{\tau_{APE,zd|z=0}}.$$

In other words, under the additional restriction that $\mathrm{V}[r(X) \mid z=1] = \mathrm{V}[r(X) \mid z=0]$, we can only expect τ_{iv} to recover $\tau_{APE,zy/zd}$ if either both instrument levels are received by equal-sized subpopulations or the effects for groups with $z=1$ and $z=0$ are the same. What follows, different instruments, corresponding to different values of π, might be more or less likely to allow τ_{iv} to recover the local average treatment effect.

5 Conclusion

In this paper I study the interpretation of the ordinary and two-stage least squares estimands in the homogeneous linear model when treatment effects are in fact heterogeneous. This problem is highly relevant for applied researchers who often rely on these simple estimation methods to provide estimates of the effects of various treatments, even though treatment effect heterogeneity is often empirically important. How should we interpret the estimates in these studies? I derive a new theoretical result which demonstrates that both ordinary and two-stage least squares estimands are equivalent to convex

[20] See, *e.g.*, Frölich (2007) for a general expression for the local average treatment effect with covariates, of which $\tau_{APE,zy/zd}$ is a special case.

combinations of two other parameters (different for OLS and 2SLS), which can be interpreted as the average treatment effect on the treated and the average treatment effect on the controls under additional assumptions. Perhaps surprisingly, the weight which is placed by OLS and 2SLS on the average effect on each group (treated or controls) is inversely related to the proportion of this group. The more units get treatment, the less weight is placed on the effect on the treated.

A pessimistic conclusion might be that OLS or 2SLS estimation of the model in (1) is inappropriate in the presence of treatment effect heterogeneity. However, it is also possible to present a more pragmatic view of my main result. Indeed, in this paper I derive a number of corollaries of this result which, in turn, lead to several diagnostic methods that I recommend to applied researchers. In general, I assume that the researcher is ultimately interested in ATE, ATT, or both, and that she wishes to estimate the model in (1) using OLS or 2SLS but is concerned about treatment effect heterogeneity. In this case, my diagnostics are able to detect deviations of the OLS/2SLS weights from the pattern which would be necessary to consistently estimate a given parameter. Importantly, these diagnostics are very easy to implement and interpret; they are bounded between zero and one in absolute value and they give the proportion of a particular measure of heterogeneity which contributes to bias. Thus, if a given diagnostic is close to zero, OLS or 2SLS is likely a reasonable choice; but if a diagnostic is far from zero, other methods should be used. In an important special case, these diagnostics become particularly simple and immediate to report. If our goal is to estimate ATT, we should simply report $\hat{\rho}$, the sample proportion of treated units; if we wish to estimate ATE, it is instead useful to report $\hat{\delta} = 2\hat{\rho} - 1$. In short, OLS and 2SLS are expected to provide a reasonable approximation to ATE if both groups, treated and controls, are of similar size. If we wish to estimate ATT, it is necessary that the proportion of treated units is quite small.

A related issue is that of construction of control groups in advance of the empirical analysis. Often the size of the treated group is essentially fixed, but the number of control units is easier to manipulate. My main result has important implications for determining the size of the control group in this context. If we choose ATT as our target parameter, the number of control units should be as large as possible, and at least several times larger than the number of treated units. An example is given by the NSW–CPS and NSW–PSID samples, analyzed by LaLonde (1986), Dehejia and Wahba (1999), Smith and Todd (2005), and many others. If instead we choose to estimate ATE, the sample proportion of control units should ideally be equal to the population proportion of treated units.

Future work might consider various extensions of my results, and their usage in testing is one possible avenue. My weighted least squares estimates (Theorem 2) could be used in a formal comparison with OLS as a specification test in the spirit of White (1980). Also, similar to Lochner and Moretti (2015), it seems possible to construct an exogeneity test which would reweight the OLS estimates of ATT and ATC using the 2SLS weights and compare the result of this reweighting with the 2SLS estimate.

References

ABADIE, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

——— (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72, 1–19.

ABADIE, A., S. ATHEY, G. W. IMBENS, AND J. M. WOOLDRIDGE (2017): "Sampling-based vs. Design-based Uncertainty in Regression Analysis." Unpublished.

ABRAHAM, S. AND L. SUN (2018): "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." Unpublished.

AIZER, A., S. ELI, J. FERRIE, AND A. LLERAS-MUNEY (2016): "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, 106, 935–971.

ALESINA, A., P. GIULIANO, AND N. NUNN (2013): "On the Origins of Gender Roles: Women and the Plough," *Quarterly Journal of Economics*, 128, 469–530.

ALMOND, D., K. Y. CHAY, AND D. S. LEE (2005): "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, 120, 1031–1083.

ALTONJI, J. G., T. E. ELDER, AND C. R. TABER (2005): "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *Journal of Human Resources*, 40, 791–821.

ANDREWS, I. (2017): "On the Structure of IV Estimands." Unpublished.

ANGRIST, J. D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249–288.

ANGRIST, J. D., K. GRADDY, AND G. W. IMBENS (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67, 499–527.

ANGRIST, J. D. AND G. W. IMBENS (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

ANGRIST, J. D., G. W. IMBENS, AND D. B. RUBIN (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

ANGRIST, J. D. AND A. B. KRUEGER (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, North-Holland, vol. 3A.

ANGRIST, J. D. AND J.-S. PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

ARONOW, P. M. AND C. SAMII (2016): "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science*, 60, 250–267.

ATHEY, S. AND G. W. IMBENS (2018): "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." Unpublished.

ATKIN, D. (2016): "The Caloric Costs of Culture: Evidence from Indian Migrants," *American Economic Review*, 106, 1144–1181.

BELENZON, S., A. K. CHATTERJI, AND B. DALEY (2017): "Eponymous Entrepreneurs," *American Economic Review*, 107, 1638–1655.

BERGER, D., W. EASTERLY, N. NUNN, AND S. SATYANATH (2013): "Commercial Imperialism? Political Influence and Trade During the Cold War," *American Economic Review*, 103, 863–896.

BETTINGER, E. P., L. FOX, S. LOEB, AND E. S. TAYLOR (2017): "Virtual Classrooms: How Online College Courses Affect Student Success," *American Economic Review*, 107, 2855–2875.

BITLER, M. P., J. B. GELBACH, AND H. W. HOYNES (2006): "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review*, 96, 988–1012.

——— (2008): "Distributional Impacts of the Self-Sufficiency Project," *Journal of Public Economics*, 92, 748–765.

BLACK, D. A., S. G. SANDERS, E. J. TAYLOR, AND L. J. TAYLOR (2015): "The Impact of the Great Migration on Mortality of African Americans: Evidence from the Deep South," *American Economic Review*, 105, 477–503.

BLACK, D. A., J. A. SMITH, M. C. BERGER, AND B. J. NOEL (2003): "Is the Threat of Reemployment Services More Effective Than the Services Themselves? Evidence from Random Assignment in the UI System," *American Economic Review*, 93, 1313–1327.

BLINDER, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.

BLUNDELL, R. AND R. L. MATZKIN (2014): "Control Functions in Nonseparable Simultaneous Equations Models," *Quantitative Economics*, 5, 271–295.

BOBONIS, G. J., L. R. CÁMARA FUERTES, AND R. SCHWABE (2016): "Monitoring Corruptible Politicians," *American Economic Review*, 106, 2371–2405.

BORUSYAK, K. AND X. JARAVEL (2017): "Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume." Unpublished.

BRINCH, C. N., M. MOGSTAD, AND M. WISWALL (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039.

BUGNI, F. A., I. A. CANAY, AND A. M. SHAIKH (2017): "Inference under Covariate-Adaptive Randomization," *Journal of the American Statistical Association*, forthcoming.

CALLAWAY, B. AND P. H. C. SANT'ANNA (2018): "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment." Unpublished.

CAMPBELL, J. Y., S. GIGLIO, AND P. PATHAK (2011): "Forced Sales and House Prices," *American Economic Review*, 101, 2108–2131.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, J. HAHN, AND W. NEWEY (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 535–580.

CLARK, D. AND E. DEL BONO (2016): "The Long-Run Effects of Attending an Elite School: Evidence from the United Kingdom," *American Economic Journal: Applied Economics*, 8, 150–176.

DAS, J., A. HOLLA, A. MOHPAL, AND K. MURALIDHARAN (2016): "Quality and Accountability in Health Care Delivery: Audit-Study Evidence from Primary Care in India," *American Economic Review*, 106, 3765–3799.

DE CHAISEMARTIN, C. AND X. D'HAULTFŒUILLE (2018): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." Unpublished.

DEATON, A. (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Johns Hopkins University Press.

DEHEJIA, R. H. AND S. WAHBA (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

DEMING, D. J., J. S. HASTINGS, T. J. KANE, AND D. O. STAIGER (2014): "School Choice, School Quality, and Postsecondary Attainment," *American Economic Review*, 104, 991–1013.

DINKELMAN, T. (2011): "The Effects of Rural Electrification on Employment: New Evidence from South Africa," *American Economic Review*, 101, 3078–3108.

DITTMAR, J. E. (2011): "Information Technology and Economic Change: The Impact of the Printing Press," *Quarterly Journal of Economics*, 126, 1133–1172.

DOBBIE, W. AND J. SONG (2015): "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection," *American Economic Review*, 105, 1272–1311.

ELDER, T. E., J. H. GODDEERIS, AND S. J. HAIDER (2010): "Unexplained Gaps and Oaxaca–Blinder Decompositions," *Labour Economics*, 17, 284–290.

EVDOKIMOV, K. S. AND M. KOLESÁR (2018): "Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects." Unpublished.

FORTIN, N., T. LEMIEUX, AND S. FIRPO (2011): "Decomposition Methods in Economics," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, North-Holland, vol. 4A.

FRAKES, M. (2013): "The Impact of Medical Liability Standards on Regional Variations in Physician Behavior: Evidence from the Adoption of National-Standard Rules," *American Economic Review*, 103, 257–276.

FREEDMAN, D. A. (2008a): "On Regression Adjustments in Experiments with Several Treatments," *Annals of Applied Statistics*, 2, 176–196.

——— (2008b): "On Regression Adjustments to Experimental Data," *Advances in Applied Mathematics*, 40, 180–193.

FRISCH, R. AND F. V. WAUGH (1933): "Partial Time Regressions as Compared with Individual Trends," *Econometrica*, 1, 387–401.

FRÖLICH, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

GIBBONS, C. E., J. C. SUÁREZ SERRATO, AND M. B. URBANCIC (2018): "Broken or Fixed Effects?" *Journal of Econometric Methods*, forthcoming.

GOODMAN-BACON, A. (2018): "Difference-in-Differences with Variation in Treatment Timing." Unpublished.

HAUSMAN, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

HECKMAN, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46, 931–959.

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

——— (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748.

HECKMAN, J. J., H. ICHIMURA, J. A. SMITH, AND P. E. TODD (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

HECKMAN, J. J., J. L. TOBIAS, AND E. J. VYTLACIL (2001): "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal*, 68, 210–223.

——— (2003): "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85, 748–755.

HECKMAN, J. J., S. URZUA, AND E. J. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432.

HECKMAN, J. J. AND E. J. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

——— (2007): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, North-Holland, vol. 6B.

HOLLAND, P. W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

HULL, P. (2018): "Estimating Treatment Effects in Mover Designs." Unpublished.

HUMPHREYS, M. (2009): "Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogeneous Assignment Probabilities." Unpublished.

IMAI, K. AND I. S. KIM (2017): "When Should We Use Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" Unpublished.

IMBENS, G. W. (2015): "Matching Methods in Practice: Three Examples," *Journal of Human Resources*, 50, 373–419.

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

IMBENS, G. W. AND D. B. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

IMBENS, G. W. AND J. M. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

JACOB, B. A. AND J. LUDWIG (2012): "The Effects of Housing Assistance on Labor Supply: Evidence from a Voucher Lottery," *American Economic Review*, 102, 272–304.

KATO, R. AND Y. SASAKI (2017): "On Using Linear Quantile Regressions for Causal Inference," *Econometric Theory*, 33, 664–690.

KHWAJA, A., G. PICONE, M. SALM, AND J. G. TROGDON (2011): "A Comparison of Treatment Effects Estimators Using a Structural Model of AMI Treatment Choices and Severity of Illness Information from Hospital Charts," *Journal of Applied Econometrics*, 26, 825–853.

KLINE, P. (2011): "Oaxaca-Blinder as a Reweighting Estimator," *American Economic Review: Papers & Proceedings*, 101, 532–537.

——— (2014): "A Note on Variance Estimation for the Oaxaca Estimator of Average Treatment Effects," *Economics Letters*, 122, 428–431.

KLINE, P. M. AND C. R. WALTERS (2018): "On Heckits, LATE, and Numerical Equivalence," Unpublished.

KOLESÁR, M. (2013): "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity." Unpublished.

LALONDE, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

LIN, W. (2013): "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *Annals of Applied Statistics*, 7, 295–318.

LØKEN, K. V., M. MOGSTAD, AND M. WISWALL (2012): "What Linear Estimators Miss: The Effects of Family Income on Child Outcomes," *American Economic Journal: Applied Economics*, 4, 1–35.

LOCHNER, L. AND E. MORETTI (2015): "Estimating and Testing Models with Many Treatment Levels and Limited Instruments," *Review of Economics and Statistics*, 97, 387–397.

LUNDBORG, P., E. PLUG, AND A. W. RASMUSSEN (2017): "Can Women Have Children and a Career? IV Evidence from IVF Treatments," *American Economic Review*, 107, 1611–1637.

MAESTAS, N., K. J. MULLEN, AND A. STRAND (2013): "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *American Economic Review*, 103, 1797–1829.

MARTINEZ-BRAVO, M. (2014): "The Role of Local Officials in New Democracies: Evidence from Indonesia," *American Economic Review*, 104, 1244–1287.

MICHALOPOULOS, S. AND E. PAPAIOANNOU (2016): "The Long-Run Effects of the Scramble for Africa," *American Economic Review*, 106, 1802–1848.

MOFFITT, R. (2008): "Estimating Marginal Treatment Effects in Heterogeneous Populations," *Annales d'Économie et de Statistique*, 91/92, 239–261.

MOSER, P., A. VOENA, AND F. WALDINGER (2014): "German Jewish Émigrés and US Invention," *American Economic Review*, 104, 3222–3255.

OAXACA, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693–709.

OLSEN, R. J. (1980): "A Least Squares Correction for Selectivity Bias," *Econometrica*, 48, 1815–1820.

PAREY, M. AND F. WALDINGER (2011): "Studying Abroad and the Effect on International Labour Market Mobility: Evidence from the Introduction of ERASMUS," *Economic Journal*, 121, 194–222.

RHODES, W. (2010): "Heterogeneous Treatment Effects: What Does a Regression Estimate?" *Evaluation Review*, 34, 334–361.

SMITH, J. A. AND P. E. TODD (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125, 305–353.

SOLON, G., S. J. HAIDER, AND J. M. WOOLDRIDGE (2015): "What Are We Weighting For?" *Journal of Human Resources*, 50, 301–316.

STREZHNEV, A. (2018): "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." Unpublished.

TELSER, L. G. (1964): "Iterative Estimation of a Set of Linear Regression Equations," *Journal of the American Statistical Association*, 59, 845–862.

VOIGTLÄNDER, N. AND H.-J. VOTH (2012): "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany," *Quarterly Journal of Economics*, 127, 1339–1392.

WHITE, H. (1980): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149–170.

WOOLDRIDGE, J. M. (2005): "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *Review of Economics and Statistics*, 87, 385–390.

——— (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2nd ed.

——— (2015): "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, 50, 420–445.

YITZHAKI, S. (1996): "On Using Linear Regressions in Welfare Economics," *Journal of Business & Economic Statistics*, 14, 478–486.

For Online Publication

A Proof of Theorem 1

First, consider equation (4), $\mathrm{L}\left(y \mid 1, d, X\right) = \alpha + \tau d + X\beta$. By the Frisch–Waugh theorem, $\tau = \tau_a$, where τ_a is defined by

$$\mathrm{L}\left[y \mid 1, d, p\left(X\right)\right] = \alpha_a + \tau_a d + \gamma_a \cdot p\left(X\right). \tag{44}$$

Second, notice that (44) is a linear projection of y on two variables: one binary, d, and one arbitrarily discrete or continuous, $p\left(X\right)$. We can therefore use the following result from Elder, Goddeeris, and Haider (2010):

Lemma 1 (Elder *et al.*, 2010) *Let* $\mathrm{L}\left(y \mid 1, d, x\right) = \alpha_e + \tau_e d + \beta_e x$ *denote the linear projection of* y *on* d *(a binary variable) and* x *(a single, possibly continuous, control variable) and let* $\mathrm{V}\left(\cdot\right)$, $\mathrm{Cov}\left(\cdot\right)$, $\mathrm{V}\left(\cdot \mid \cdot\right)$, *and* $\mathrm{Cov}\left(\cdot \mid \cdot\right)$ *denote the variance, the covariance, the conditional variance, and the conditional covariance, respectively. Then,*

$$\begin{aligned} \tau_e &= \frac{\rho \cdot \mathrm{V}\left(x \mid d=1\right)}{\rho \cdot \mathrm{V}\left(x \mid d=1\right) + \left(1-\rho\right) \cdot \mathrm{V}\left(x \mid d=0\right)} \cdot \theta_1 \\ &+ \frac{\left(1-\rho\right) \cdot \mathrm{V}\left(x \mid d=0\right)}{\rho \cdot \mathrm{V}\left(x \mid d=1\right) + \left(1-\rho\right) \cdot \mathrm{V}\left(x \mid d=0\right)} \cdot \theta_0, \end{aligned}$$

where

$$\theta_1 = \frac{\mathrm{Cov}\left(d, y\right)}{\mathrm{V}\left(d\right)} - \frac{\mathrm{Cov}\left(d, x\right)}{\mathrm{V}\left(d\right)} \cdot \frac{\mathrm{Cov}\left(x, y \mid d=1\right)}{\mathrm{V}\left(x \mid d=1\right)}$$

and

$$\theta_0 = \frac{\mathrm{Cov}\left(d, y\right)}{\mathrm{V}\left(d\right)} - \frac{\mathrm{Cov}\left(d, x\right)}{\mathrm{V}\left(d\right)} \cdot \frac{\mathrm{Cov}\left(x, y \mid d=0\right)}{\mathrm{V}\left(x \mid d=0\right)}.$$

Combining the two pieces gives

$$\begin{aligned} \tau &= \frac{\rho \cdot \mathrm{V}\left[p\left(X\right) \mid d=1\right]}{\rho \cdot \mathrm{V}\left[p\left(X\right) \mid d=1\right] + \left(1-\rho\right) \cdot \mathrm{V}\left[p\left(X\right) \mid d=0\right]} \cdot \theta_1^* \\ &+ \frac{\left(1-\rho\right) \cdot \mathrm{V}\left[p\left(X\right) \mid d=0\right]}{\rho \cdot \mathrm{V}\left[p\left(X\right) \mid d=1\right] + \left(1-\rho\right) \cdot \mathrm{V}\left[p\left(X\right) \mid d=0\right]} \cdot \theta_0^*, \end{aligned} \tag{45}$$

where

$$\theta_1^* = \frac{\mathrm{Cov}\left(d, y\right)}{\mathrm{V}\left(d\right)} - \frac{\mathrm{Cov}\left[d, p\left(X\right)\right]}{\mathrm{V}\left(d\right)} \cdot \frac{\mathrm{Cov}\left[p\left(X\right), y \mid d=1\right]}{\mathrm{V}\left[p\left(X\right) \mid d=1\right]} \tag{46}$$

and

$$\theta_0^* = \frac{\mathrm{Cov}\,(d, y)}{\mathrm{V}\,(d)} - \frac{\mathrm{Cov}\,[d, p\,(X)]}{\mathrm{V}\,(d)} \cdot \frac{\mathrm{Cov}\,[p\,(X)\,, y \mid d = 0]}{\mathrm{V}\,[p\,(X) \mid d = 0]}. \tag{47}$$

Third, notice that $\theta_1^* = \tau_{APE|d=0}$ and $\theta_0^* = \tau_{APE|d=1}$, as defined in (11). Indeed,

$$\frac{\mathrm{Cov}\,(d, y)}{\mathrm{V}\,(d)} = \mathrm{E}\,(y \mid d = 1) - \mathrm{E}\,(y \mid d = 0) \tag{48}$$

and also

$$\frac{\mathrm{Cov}\,[d, p\,(X)]}{\mathrm{V}\,(d)} = \mathrm{E}\,[p\,(X) \mid d = 1] - \mathrm{E}\,[p\,(X) \mid d = 0]\,. \tag{49}$$

Moreover, for $j = 0, 1$,

$$\frac{\mathrm{Cov}\,[p\,(X)\,, y \mid d = j]}{\mathrm{V}\,[p\,(X) \mid d = j]} = \gamma_j \tag{50}$$

where γ_1 and γ_0 are defined in (8) and (9), respectively. Because

$$\begin{aligned} \mathrm{E}\,(y \mid d = 1) - \mathrm{E}\,(y \mid d = 0) &= \{\mathrm{E}\,[p\,(X) \mid d = 1] - \mathrm{E}\,[p\,(X) \mid d = 0]\} \cdot \gamma_1 \\ &+ (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot \mathrm{E}\,[p\,(X) \mid d = 0] \end{aligned} \tag{51}$$

and also

$$\begin{aligned} \mathrm{E}\,(y \mid d = 1) - \mathrm{E}\,(y \mid d = 0) &= \{\mathrm{E}\,[p\,(X) \mid d = 1] - \mathrm{E}\,[p\,(X) \mid d = 0]\} \cdot \gamma_0 \\ &+ (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot \mathrm{E}\,[p\,(X) \mid d = 1]\,, \end{aligned} \tag{52}$$

where again α_1 and α_0 are defined in (8) and (9), we get the result that $\theta_1^* = \tau_{APE|d=0}$ and $\theta_0^* = \tau_{APE|d=1}$. Note that equations (51) and (52) are special cases of the Oaxaca–Blinder decomposition (Blinder, 1973; Oaxaca, 1973; Fortin, Lemieux, and Firpo, 2011). Finally, combining the three pieces gives

$$\begin{aligned} \tau &= \frac{\rho \cdot \mathrm{V}\,[p\,(X) \mid d = 1]}{\rho \cdot \mathrm{V}\,[p\,(X) \mid d = 1] + (1 - \rho) \cdot \mathrm{V}\,[p\,(X) \mid d = 0]} \cdot \tau_{APE|d=0} \\ &+ \frac{(1 - \rho) \cdot \mathrm{V}\,[p\,(X) \mid d = 0]}{\rho \cdot \mathrm{V}\,[p\,(X) \mid d = 1] + (1 - \rho) \cdot \mathrm{V}\,[p\,(X) \mid d = 0]} \cdot \tau_{APE|d=1}, \end{aligned} \tag{53}$$

which completes the proof. □

B Proportion of Treated Units and OLS Weights

To show formally that w_1 is decreasing in ρ and that w_0 is increasing in ρ, it is convenient to additionally assume that $\mathrm{E}(d \mid X)$ is linear in X. This restriction arises naturally in the analysis of saturated models by Angrist (1998) and Humphreys (2009). It is also used by Rhodes (2010), Aronow and Samii (2016), and Abadie *et al.* (2017). In the present context there are two reasons why this linearity assumption is useful. First, it allows us to rewrite w_0 and w_1 solely in terms of unconditional expectations of $p(X)$ and its powers. Second, it simplifies calculation of the derivatives of w_0 and w_1 with respect to the intercept of the propensity score model. Imposing a shift on this intercept is equivalent to changing ρ by a small amount. It turns out that Theorem 1 and this additional linearity assumption imply the following result.

Corollary 6 *Suppose that* $\mathrm{E}(d \mid X) = p(X) = \alpha_p + X\beta_p$. *Then, Theorem 1 implies that*

$$\frac{dw_1}{d\alpha_p} < 0 \qquad \textit{and} \qquad \frac{dw_0}{d\alpha_p} > 0.$$

Proof of Corollary 6. For simplicity, we first focus on a_0 and a_1, which we define as $a_0 = \rho \cdot \mathrm{V}[p(X) \mid d=1]$ and $a_1 = (1-\rho) \cdot \mathrm{V}[p(X) \mid d=0]$. What follows, $w_0 = \frac{a_0}{a_0+a_1}$ and $w_1 = \frac{a_1}{a_0+a_1}$. It turns out that we can rewrite a_0 as

$$\begin{aligned}
a_0 &= \mathrm{E}(d) \cdot \mathrm{E}\left(\{p(X) - \mathrm{E}[p(X) \mid d=1]\}^2 \mid d=1\right) \\
&= \mathrm{E}(d) \cdot \left(\mathrm{E}\left[p(X)^2 \mid d=1\right] - \{\mathrm{E}[p(X) \mid d=1]\}^2\right) \\
&= \mathrm{E}(d) \cdot \left(\frac{\mathrm{E}\left[p(X)^2 d\right]}{\mathrm{E}(d)} - \left\{\frac{\mathrm{E}[p(X)d]}{\mathrm{E}(d)}\right\}^2\right) \\
&= \mathrm{E}\left[p(X)^2 d\right] - \frac{\{\mathrm{E}[p(X)d]\}^2}{\mathrm{E}(d)} \\
&= \mathrm{E}\left[p(X)^2 \mathrm{E}(d \mid X)\right] - \frac{\{\mathrm{E}[p(X)\mathrm{E}(d \mid X)]\}^2}{\mathrm{E}[\mathrm{E}(d \mid X)]} \\
&= \mathrm{E}\left[p(X)^3\right] - \frac{\left\{\mathrm{E}\left[p(X)^2\right]\right\}^2}{\mathrm{E}[p(X)]}. \qquad (54)
\end{aligned}$$

Then, taking the derivative of a_0 with respect to α_p gives

$$\begin{aligned}
\frac{da_0}{d\alpha_p} &= 3\mathrm{E}\left[p\left(X\right)^2\right] - \frac{4\mathrm{E}\left[p\left(X\right)^2\right]\mathrm{E}\left[p\left(X\right)\right]}{\mathrm{E}\left[p\left(X\right)\right]} + \frac{\left\{\mathrm{E}\left[p\left(X\right)^2\right]\right\}^2}{\mathrm{E}\left[p\left(X\right)\right]^2} \\
&= -\mathrm{E}\left[p\left(X\right)^2\right] + \frac{\left\{\mathrm{E}\left[p\left(X\right)^2\right]\right\}^2}{\mathrm{E}\left[p\left(X\right)\right]^2} \\
&= \frac{\left\{\mathrm{E}\left[p\left(X\right)^2\right]\right\}^2 - \mathrm{E}\left[p\left(X\right)^2\right]\mathrm{E}\left[p\left(X\right)\right]^2}{\mathrm{E}\left[p\left(X\right)\right]^2} \\
&= \frac{\mathrm{E}\left[p\left(X\right)^2\right]\left\{\mathrm{E}\left[p\left(X\right)^2\right] - \mathrm{E}\left[p\left(X\right)\right]^2\right\}}{\mathrm{E}\left[p\left(X\right)\right]^2} \\
&= \frac{\mathrm{E}\left[p\left(X\right)^2\right]\mathrm{V}\left[p\left(X\right)\right]}{\mathrm{E}\left[p\left(X\right)\right]^2} > 0. \qquad (55)
\end{aligned}$$

Similarly,

$$\begin{aligned}
a_1 &= \left[1-\mathrm{E}\left(d\right)\right]\cdot\mathrm{E}\left(\left\{p\left(X\right)-\mathrm{E}\left[p\left(X\right)\mid d=0\right]\right\}^2 \mid d=0\right) \\
&= \left[1-\mathrm{E}\left(d\right)\right]\cdot\left(\mathrm{E}\left[p\left(X\right)^2 \mid d=0\right] - \left\{\mathrm{E}\left[p\left(X\right)\mid d=0\right]\right\}^2\right) \\
&= \left[1-\mathrm{E}\left(d\right)\right]\cdot\left(\frac{\mathrm{E}\left[p\left(X\right)^2\right]-\mathrm{E}\left[p\left(X\right)^2 d\right]}{1-\mathrm{E}\left(d\right)} - \left\{\frac{\mathrm{E}\left[p\left(X\right)\right]-\mathrm{E}\left[p\left(X\right)d\right]}{1-\mathrm{E}\left(d\right)}\right\}^2\right) \\
&= \mathrm{E}\left[p\left(X\right)^2\right]-\mathrm{E}\left[p\left(X\right)^2 d\right] - \frac{\left\{\mathrm{E}\left[p\left(X\right)\right]-\mathrm{E}\left[p\left(X\right)d\right]\right\}^2}{1-\mathrm{E}\left(d\right)} \\
&= \mathrm{E}\left[p\left(X\right)^2\right]-\mathrm{E}\left[p\left(X\right)^2\mathrm{E}\left(d\mid X\right)\right] - \frac{\left\{\mathrm{E}\left[p\left(X\right)\right]-\mathrm{E}\left[p\left(X\right)\mathrm{E}\left(d\mid X\right)\right]\right\}^2}{1-\mathrm{E}\left[\mathrm{E}\left(d\mid X\right)\right]} \\
&= \mathrm{E}\left[p\left(X\right)^2\right]-\mathrm{E}\left[p\left(X\right)^3\right] - \frac{\left\{\mathrm{E}\left[p\left(X\right)\right]-\mathrm{E}\left[p\left(X\right)^2\right]\right\}^2}{1-\mathrm{E}\left[p\left(X\right)\right]} \qquad (56)
\end{aligned}$$

and

$$\begin{aligned}
\frac{da_1}{d\alpha_p} &= 2\mathrm{E}\left[p\left(X\right)\right] - 3\mathrm{E}\left[p\left(X\right)^2\right] - \frac{\left\{\mathrm{E}\left[p\left(X\right)\right]-\mathrm{E}\left[p\left(X\right)^2\right]\right\}^2}{\left\{1-\mathrm{E}\left[p\left(X\right)\right]\right\}^2} \\
&- \frac{2\cdot\left\{1-\mathrm{E}\left[p\left(X\right)\right]\right\}\cdot\left\{1-2\mathrm{E}\left[p\left(X\right)\right]\right\}\cdot\left\{\mathrm{E}\left[p\left(X\right)\right]-\mathrm{E}\left[p\left(X\right)^2\right]\right\}}{\left\{1-\mathrm{E}\left[p\left(X\right)\right]\right\}^2}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathrm{E}\left[p\left(X\right)\right]^2 - \mathrm{E}\left[p\left(X\right)^2\right]}{\left\{1-\mathrm{E}\left[p\left(X\right)\right]\right\}^2} \\
&+ \frac{2\mathrm{E}\left[p\left(X\right)\right]\mathrm{E}\left[p\left(X\right)^2\right] - 2\mathrm{E}\left[p\left(X\right)\right]^3}{\left\{1-\mathrm{E}\left[p\left(X\right)\right]\right\}^2} \\
&+ \frac{\mathrm{E}\left[p\left(X\right)^2\right]\mathrm{E}\left[p\left(X\right)\right]^2 - \left\{\mathrm{E}\left[p\left(X\right)^2\right]\right\}^2}{\left\{1-\mathrm{E}\left[p\left(X\right)\right]\right\}^2} \\
&= \frac{-\mathrm{V}\left[p\left(X\right)\right]\cdot\left\{1-2\mathrm{E}\left[p\left(X\right)\right]+\mathrm{E}\left[p\left(X\right)^2\right]\right\}}{\left\{1-\mathrm{E}\left[p\left(X\right)\right]\right\}^2} \\
&= \frac{-\mathrm{V}\left[p\left(X\right)\right]\cdot\mathrm{E}\left\{\left[1-p\left(X\right)\right]^2\right\}}{\left\{1-\mathrm{E}\left[p\left(X\right)\right]\right\}^2} < 0. \qquad (57)
\end{aligned}
$$

Finally, it follows that

$$\frac{dw_1}{d\alpha_p} < 0 \quad \text{and} \quad \frac{dw_0}{d\alpha_p} > 0, \qquad (58)$$

since $w_0 = \frac{a_0}{a_0+a_1}$, $w_1 = \frac{a_1}{a_0+a_1}$, $a_0 > 0$, $a_1 > 0$, $\frac{da_0}{d\alpha_p} > 0$, and $\frac{da_1}{d\alpha_p} < 0$. □

C Implicit Residualization

To gain further intuition for Theorem 1 it is useful to relate τ, the ordinary least squares estimand, to partial residualization that is implicit in least squares estimation. In other words, after projecting y on d and X, we can subtract $X\beta$ from y. Since the coefficient on d will be the same in the linear projection of the new variable on d, we might get further insight into τ by studying the interpretation of this new variable—the partially residualized y. This derivation also leads to an alternative proof of Theorem 1.

Proof of Theorem 1. As before, consider equation (4), $\mathrm{L}\left(y \mid 1, d, X\right) = \alpha + \tau d + X\beta$, and note that $\tau = \tau_a$, where τ_a is defined by $\mathrm{L}\left[y \mid 1, d, p\left(X\right)\right] = \alpha_a + \tau_a d + \gamma_a \cdot p\left(X\right)$. We can write this linear projection in error form as

$$y = \alpha_a + \tau_a d + \gamma_a \cdot p\left(X\right) + v. \tag{59}$$

We also consider separate linear projections for $d = 1$ and $d = 0$, namely

$$\mathrm{L}\left[y \mid 1, p\left(X\right)\right] = \alpha_1 + \gamma_1 \cdot p\left(X\right) \quad \text{if} \quad d = 1 \tag{60}$$

and

$$\mathrm{L}\left[y \mid 1, p\left(X\right)\right] = \alpha_0 + \gamma_0 \cdot p\left(X\right) \quad \text{if} \quad d = 0. \tag{61}$$

Henceforth, to simplify notation I will use $l_1(X)$ to denote $\alpha_1 + \gamma_1 \cdot p\left(X\right)$ and $l_0(X)$ to denote $\alpha_0 + \gamma_0 \cdot p\left(X\right)$. To understand the relationship between γ_a, γ_1, and γ_0, we can use the following result from Deaton (1997) and Solon *et al.* (2015):

Lemma 2 (Deaton, 1997; Solon *et al.*, 2015) *Let* $\mathrm{L}\left(y \mid 1, d, x\right) = \alpha_e + \tau_e d + \beta_e x$ *denote the linear projection of y on d (binary) and x (possibly continuous). Then,*

$$\begin{aligned} \beta_e &= \frac{\rho \cdot \mathrm{V}\left(x \mid d = 1\right)}{\rho \cdot \mathrm{V}\left(x \mid d = 1\right) + \left(1 - \rho\right) \cdot \mathrm{V}\left(x \mid d = 0\right)} \cdot \beta_{1,e} \\ &+ \frac{\left(1 - \rho\right) \cdot \mathrm{V}\left(x \mid d = 0\right)}{\rho \cdot \mathrm{V}\left(x \mid d = 1\right) + \left(1 - \rho\right) \cdot \mathrm{V}\left(x \mid d = 0\right)} \cdot \beta_{0,e}, \end{aligned}$$

where $\beta_{1,e}$ *and* $\beta_{0,e}$ *are defined by*

$$\mathrm{L}\left(y \mid 1, x\right) = \alpha_{1,e} + \beta_{1,e} x \quad \textit{if} \quad d = 1$$

and

$$\mathrm{L}\left(y \mid 1, x\right) = \alpha_{0,e} + \beta_{0,e} x \quad \textit{if} \quad d = 0.$$

An implication of Lemma 2 is that

$$\gamma_a = w_0 \cdot \gamma_1 + w_1 \cdot \gamma_0. \tag{62}$$

Next, we can rewrite equation (59) as

$$\begin{aligned} y - w_0 \cdot \gamma_1 \cdot p\left(X\right) - w_1 \cdot \gamma_0 \cdot p\left(X\right) &= \alpha_a + \tau_a d + v \\ &= \mathrm{E}\left(y\right) - \tau_a \cdot \mathrm{E}\left(d\right) - \gamma_a \cdot \mathrm{E}\left[p\left(X\right)\right] \\ &+ \tau_a d + v. \end{aligned} \tag{63}$$

Moreover, it turns out that

$$\alpha_1 = \mathrm{E}\left(y \mid d = 1\right) - \gamma_1 \cdot \mathrm{E}\left[p\left(X\right) \mid d = 1\right] \tag{64}$$

and also

$$\alpha_0 = \mathrm{E}\left(y \mid d = 0\right) - \gamma_0 \cdot \mathrm{E}\left[p\left(X\right) \mid d = 0\right]. \tag{65}$$

What follows,

$$\begin{aligned} y - w_0 \cdot l_1(X) - w_1 \cdot l_0(X) &= \mathrm{E}\left(y\right) - w_0 \cdot \mathrm{E}\left(y \mid d = 1\right) - w_1 \cdot \mathrm{E}\left(y \mid d = 0\right) \\ &+ w_0 \cdot \gamma_1 \cdot \left\{\mathrm{E}\left[p\left(X\right) \mid d = 1\right] - \mathrm{E}\left[p\left(X\right)\right]\right\} \\ &+ w_1 \cdot \gamma_0 \cdot \left\{\mathrm{E}\left[p\left(X\right) \mid d = 0\right] - \mathrm{E}\left[p\left(X\right)\right]\right\} \\ &- \tau_a \cdot \mathrm{E}\left(d\right) + \tau_a d + v. \end{aligned} \tag{66}$$

In other words, in a linear projection of $y - w_0 \cdot l_1(X) - w_1 \cdot l_0(X)$ on d, the coefficient on d is equal to τ_a and the intercept is equal to $\mathrm{E}\left(y\right) - w_0 \cdot \mathrm{E}\left(y \mid d = 1\right) - w_1 \cdot \mathrm{E}\left(y \mid d = 0\right) + w_0 \cdot \gamma_1 \cdot \left\{\mathrm{E}\left[p\left(X\right) \mid d = 1\right] - \mathrm{E}\left[p\left(X\right)\right]\right\} + w_1 \cdot \gamma_0 \cdot \left\{\mathrm{E}\left[p\left(X\right) \mid d = 0\right] - \mathrm{E}\left[p\left(X\right)\right]\right\} - \tau_a \cdot \mathrm{E}\left(d\right)$. However, τ_a must also be equal to the difference in expected values of the dependent variable for $d = 1$ and $d = 0$. Using (12) and (13), we can write these expected values as

$$\mathrm{E}\left[y - w_0 \cdot l_1(X) - w_1 \cdot l_0(X) \mid d = 1\right] = w_1 \cdot \tau_{APE|d=1} \tag{67}$$

and

$$\mathrm{E}\left[y - w_0 \cdot l_1(X) - w_1 \cdot l_0(X) \mid d = 0\right] = -w_0 \cdot \tau_{APE|d=0}. \tag{68}$$

Thus,

$$\tau = \tau_a = w_1 \cdot \tau_{APE|d=1} + w_0 \cdot \tau_{APE|d=0}, \tag{69}$$

which completes the proof. □

D Comparison with the Result for Saturated Models

The baseline result in Angrist (1998) is derived for a model with two strata, where x indicates stratum membership. The result is that if $\mathrm{L}(y \mid 1, d, x) = \alpha_g + \tau_g d + \beta_g x$, then

$$\begin{aligned}\tau_g &= \frac{\mathrm{P}(x=0) \cdot \mathrm{V}(d \mid x=0)}{\mathrm{P}(x=0) \cdot \mathrm{V}(d \mid x=0) + \mathrm{P}(x=1) \cdot \mathrm{V}(d \mid x=1)} \cdot \tau_0 \\ &+ \frac{\mathrm{P}(x=1) \cdot \mathrm{V}(d \mid x=1)}{\mathrm{P}(x=0) \cdot \mathrm{V}(d \mid x=0) + \mathrm{P}(x=1) \cdot \mathrm{V}(d \mid x=1)} \cdot \tau_1, \qquad (70)\end{aligned}$$

where τ_1 and τ_0 denote the stratum-specific effects. Theorem 1 might appear at first sight to be similar to this result. There are, however, two major differences between these formulations: first, Theorem 1 conditions on d, while Angrist (1998) conditions on x, and therefore does not specify his result in terms of group-specific average partial effects; second, Angrist (1998) does not recover a pattern of "weight reversal," whose manifestation is the main result of this paper. In this appendix I show that equation (70), *i.e.* the baseline result in Angrist (1998), can be derived from a special case of Theorem 1 (or Lemma 1).

If we apply Lemma 1 to τ_g in $\mathrm{L}(y \mid 1, d, x) = \alpha_g + \tau_g d + \beta_g x$, *i.e.* to the two-strata model in Angrist (1998), we get

$$\begin{aligned}\tau_g &= \frac{\rho \cdot \mathrm{V}(x \mid d=1) \cdot [\mathrm{P}(x=0 \mid d=0) \cdot \tau_0 + \mathrm{P}(x=1 \mid d=0) \cdot \tau_1]}{\rho \cdot \mathrm{V}(x \mid d=1) + (1-\rho) \cdot \mathrm{V}(x \mid d=0)} \\ &+ \frac{(1-\rho) \cdot \mathrm{V}(x \mid d=0) \cdot [\mathrm{P}(x=0 \mid d=1) \cdot \tau_0 + \mathrm{P}(x=1 \mid d=1) \cdot \tau_1]}{\rho \cdot \mathrm{V}(x \mid d=1) + (1-\rho) \cdot \mathrm{V}(x \mid d=0)} \\ &= \frac{\rho \cdot \mathrm{V}(x \mid d=1) \cdot \mathrm{P}(x=0 \mid d=0)}{\rho \cdot \mathrm{V}(x \mid d=1) + (1-\rho) \cdot \mathrm{V}(x \mid d=0)} \cdot \tau_0 \\ &+ \frac{(1-\rho) \cdot \mathrm{V}(x \mid d=0) \cdot \mathrm{P}(x=0 \mid d=1)}{\rho \cdot \mathrm{V}(x \mid d=1) + (1-\rho) \cdot \mathrm{V}(x \mid d=0)} \cdot \tau_0 \\ &+ \frac{\rho \cdot \mathrm{V}(x \mid d=1) \cdot \mathrm{P}(x=1 \mid d=0)}{\rho \cdot \mathrm{V}(x \mid d=1) + (1-\rho) \cdot \mathrm{V}(x \mid d=0)} \cdot \tau_1 \\ &+ \frac{(1-\rho) \cdot \mathrm{V}(x \mid d=0) \cdot \mathrm{P}(x=1 \mid d=1)}{\rho \cdot \mathrm{V}(x \mid d=1) + (1-\rho) \cdot \mathrm{V}(x \mid d=0)} \cdot \tau_1, \qquad (71)\end{aligned}$$

which can be further rearranged using Bayes' theorem. Indeed,

$$\tau_g = \frac{\mathrm{P}(x=0) \cdot \mathrm{V}(d \mid x=0) \cdot \mathrm{P}(d=1 \mid x=1)}{\rho \cdot \mathrm{V}(x \mid d=1) + (1-\rho) \cdot \mathrm{V}(x \mid d=0)} \cdot \frac{\mathrm{P}(x=0) \cdot \mathrm{P}(x=1)}{\rho \cdot (1-\rho)} \cdot \tau_0$$

$$
\begin{aligned}
&+ \quad \frac{\mathrm{P}(x=0)\cdot \mathrm{V}(d\mid x=0)\cdot \mathrm{P}(d=0\mid x=1)}{\rho\cdot \mathrm{V}(x\mid d=1)+(1-\rho)\cdot \mathrm{V}(x\mid d=0)}\cdot \frac{\mathrm{P}(x=0)\cdot \mathrm{P}(x=1)}{\rho\cdot(1-\rho)}\cdot \tau_0 \\
&+ \quad \frac{\mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1)\cdot \mathrm{P}(d=1\mid x=0)}{\rho\cdot \mathrm{V}(x\mid d=1)+(1-\rho)\cdot \mathrm{V}(x\mid d=0)}\cdot \frac{\mathrm{P}(x=0)\cdot \mathrm{P}(x=1)}{\rho\cdot(1-\rho)}\cdot \tau_1 \\
&+ \quad \frac{\mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1)\cdot \mathrm{P}(d=0\mid x=0)}{\rho\cdot \mathrm{V}(x\mid d=1)+(1-\rho)\cdot \mathrm{V}(x\mid d=0)}\cdot \frac{\mathrm{P}(x=0)\cdot \mathrm{P}(x=1)}{\rho\cdot(1-\rho)}\cdot \tau_1 \\
&= \quad \frac{\mathrm{P}(x=0)\cdot \mathrm{V}(d\mid x=0)}{\rho\cdot \mathrm{V}(x\mid d=1)+(1-\rho)\cdot \mathrm{V}(x\mid d=0)}\cdot \frac{\mathrm{P}(x=0)\cdot \mathrm{P}(x=1)}{\rho\cdot(1-\rho)}\cdot \tau_0 \\
&+ \quad \frac{\mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1)}{\rho\cdot \mathrm{V}(x\mid d=1)+(1-\rho)\cdot \mathrm{V}(x\mid d=0)}\cdot \frac{\mathrm{P}(x=0)\cdot \mathrm{P}(x=1)}{\rho\cdot(1-\rho)}\cdot \tau_1 \\
&= \quad \frac{\mathrm{P}(x=0)\cdot \mathrm{V}(d\mid x=0)}{\mathrm{P}(x=0)\cdot \mathrm{V}(d\mid x=0)+\mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1)}\cdot \tau_0 \\
&+ \quad \frac{\mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1)}{\mathrm{P}(x=0)\cdot \mathrm{V}(d\mid x=0)+\mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1)}\cdot \tau_1, \qquad (72)
\end{aligned}
$$

where the last equality, again, follows from Bayes' theorem. More precisely,

$$
\begin{aligned}
\frac{\rho\cdot(1-\rho)\cdot\rho\cdot \mathrm{V}(x\mid d=1)}{\mathrm{P}(x=0)\cdot \mathrm{P}(x=1)} &= (1-\rho)\cdot \mathrm{P}(d=1\mid x=1)\cdot \mathrm{P}(d=1\mid x=0) \\
&= \mathrm{P}(x=0)\cdot \mathrm{V}(d\mid x=0)\cdot \mathrm{P}(d=1\mid x=1) \\
&+ \mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1)\cdot \mathrm{P}(d=1\mid x=0) \\
&= \lambda_1 \qquad (73)
\end{aligned}
$$

and also

$$
\begin{aligned}
\frac{\rho\cdot(1-\rho)\cdot(1-\rho)\cdot \mathrm{V}(x\mid d=0)}{\mathrm{P}(x=0)\cdot \mathrm{P}(x=1)} &= \rho\cdot \mathrm{P}(d=0\mid x=1)\cdot \mathrm{P}(d=0\mid x=0) \\
&= \mathrm{P}(x=0)\cdot \mathrm{V}(d\mid x=0)\cdot \mathrm{P}(d=0\mid x=1) \\
&+ \mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1)\cdot \mathrm{P}(d=0\mid x=0) \\
&= \lambda_0, \qquad (74)
\end{aligned}
$$

which leads to

$$
\lambda_0+\lambda_1=\mathrm{P}(x=0)\cdot \mathrm{V}(d\mid x=0)+\mathrm{P}(x=1)\cdot \mathrm{V}(d\mid x=1). \qquad (75)
$$

The equivalence between equations (70) and (72) confirms that the result in Angrist (1998) can be derived from a special case of Lemma 1, in which both d and x are binary.

E Implementation in Stata

This appendix discusses possible applications of my theoretical results in Stata. I provide separate implementations of Theorems 1, 2, 3, and 5. Additionally, I discuss how to implement my diagnostic tools for undesirable weighting of heterogeneous treatment effects in OLS and 2SLS.

Theorem 1 in Stata

Let `ovar` be the name of the outcome variable, let `tvar` be the name of the binary variable of interest ("treatment"), and let `xvars` be the list of names of other covariates. Consider the following code in Stata:

```
version 14.2
regress ovar tvar xvars
keep if e(sample)
regress tvar xvars
predict ps
regress ovar ps if tvar==1
predict ot
regress ovar ps if tvar==0
predict oc
generate te = ot-oc
summarize te if tvar==1
scalar att = r(mean)
summarize te if tvar==0
scalar atc = r(mean)
summarize ps if tvar==1
scalar v1 = r(Var)*((r(N)-1)/r(N))
summarize ps if tvar==0
scalar v0 = r(Var)*((r(N)-1)/r(N))
summarize tvar
scalar p1 = r(mean)
scalar p0 = 1-p1
scalar w1 = (p0*v0)/(p0*v0+p1*v1)
scalar w0 = (p1*v1)/(p0*v0+p1*v1)
scalar ols = att*w1+atc*w0
```

Following Theorem 1, the outcome of this procedure (`ols`) will be identical to the coefficient on `tvar` in:

```
regress ovar tvar xvars
```

Diagnostics for OLS in Stata

Let `ovar` be the name of the outcome variable, let `tvar` be the name of the binary variable of interest ("treatment"), and let `xvars` be the list of names of other covariates. Consider the following code in Stata:

```
version 14.2
regress ovar tvar xvars
keep if e(sample)
regress tvar xvars
predict ps
summarize ps if tvar==1
scalar v1 = r(Var)*((r(N)-1)/r(N))
summarize ps if tvar==0
scalar v0 = r(Var)*((r(N)-1)/r(N))
summarize tvar
scalar p1 = r(mean)
scalar p0 = 1-p1
scalar w1 = (p0*v0)/(p0*v0+p1*v1)
scalar w0 = (p1*v1)/(p0*v0+p1*v1)
scalar delta = p1-w1
```

Following Theorem 1 and Corollaries 2 and 4, w_0 and δ, my diagnostics for undesirable weighting of heterogeneous treatment effects in OLS with ATT and ATE as our target parameters, respectively, correspond to `w0` and `delta` above.

Theorem 2 in Stata

Let `ovar` be the name of the outcome variable, let `tvar` be the name of the binary variable of interest ("treatment"), and let `xvars` be the list of names of other covariates. Consider the following code in Stata:

```
version 14.2
regress ovar tvar xvars
keep if e(sample)
regress tvar xvars
predict ps
summarize ps if tvar==1
scalar v1 = r(Var)*((r(N)-1)/r(N))
summarize ps if tvar==0
scalar v0 = r(Var)*((r(N)-1)/r(N))
summarize tvar
scalar p1 = r(mean)
scalar p0 = 1-p1
scalar w1 = (p0*v0)/(p0*v0+p1*v1)
scalar w0 = (p1*v1)/(p0*v0+p1*v1)
regress ovar tvar ps [pw = tvar*(p0/w0)+(1-tvar)*(p1/w1)]
```

Following Theorem 2, the coefficient on tvar in this regression will be identical to the outcome (ate) of the following procedure:

```
regress ovar ps if tvar==1
predict ot
regress ovar ps if tvar==0
predict oc
generate te = ot-oc
summarize te
scalar ate = r(mean)
```

Theorem 3 in Stata

Let ovar be the name of the outcome variable, let tvar be the name of the binary variable of interest ("treatment"), let xvars be the list of names of other covariates, and let zvars be the list of names of instruments for tvar. Consider the following code in Stata:

```
version 14.2
ivregress 2sls ovar xvars (tvar = zvars)
keep if e(sample)
```

```
regress tvar xvars zvars
predict cf, residuals
regress tvar xvars cf
predict ps
regress ovar ps if tvar==1
predict ot
regress ovar ps if tvar==0
predict oc
generate te = ot-oc
summarize te if tvar==1
scalar att = r(mean)
summarize te if tvar==0
scalar atc = r(mean)
summarize ps if tvar==1
scalar v1 = r(Var)*((r(N)-1)/r(N))
summarize ps if tvar==0
scalar v0 = r(Var)*((r(N)-1)/r(N))
summarize tvar
scalar p1 = r(mean)
scalar p0 = 1-p1
scalar w1 = (p0*v0)/(p0*v0+p1*v1)
scalar w0 = (p1*v1)/(p0*v0+p1*v1)
scalar tsls = att*w1+atc*w0
```

Following Theorem 3, the outcome of this procedure (`tsls`) will be identical to the coefficient on `tvar` in:

```
ivregress 2sls ovar xvars (tvar = zvars)
```

Diagnostics for 2SLS in Stata

Let `ovar` be the name of the outcome variable, let `tvar` be the name of the binary variable of interest ("treatment"), let `xvars` be the list of names of other covariates, and let `zvars` be the list of names of instruments for `tvar`. Consider the following code in Stata:

```
version 14.2
```

```
ivregress 2sls ovar xvars (tvar = zvars)
keep if e(sample)
regress tvar xvars zvars
predict cf, residuals
regress tvar xvars cf
predict ps
summarize ps if tvar==1
scalar v1 = r(Var)*((r(N)-1)/r(N))
summarize ps if tvar==0
scalar v0 = r(Var)*((r(N)-1)/r(N))
summarize tvar
scalar p1 = r(mean)
scalar p0 = 1-p1
scalar w1 = (p0*v0)/(p0*v0+p1*v1)
scalar w0 = (p1*v1)/(p0*v0+p1*v1)
scalar delta = p1-w1
```

Following Theorem 3 and fn. 14, $w_{0,ts}$ and δ_{ts}, my diagnostics for undesirable weighting of heterogeneous treatment effects in 2SLS with ATT and ATE as our target parameters, respectively, correspond to `w0` and `delta` above.

Theorem 5 in Stata

Let `ovar` be the name of the outcome variable, let `tvar` be the name of the independent variable of interest ("treatment"), let `xvars` be the list of names of other covariates, and let `zvar` be the name of the binary instrument for `tvar`. Consider the following code in Stata:

```
version 14.2
ivregress 2sls ovar xvars (tvar = zvar)
keep if e(sample)
regress zvar xvars
predict ips
regress ovar ips if zvar==1
predict o1
regress ovar ips if zvar==0
predict o0
```

```
generate oe = o1-o0
summarize oe if zvar==1
scalar oe1 = r(mean)
summarize oe if zvar==0
scalar oe0 = r(mean)
regress tvar ips if zvar==1
predict t1
regress tvar ips if zvar==0
predict t0
generate te = t1-t0
summarize te if zvar==1
scalar te1 = r(mean)
summarize te if zvar==0
scalar te0 = r(mean)
summarize ips if zvar==1
scalar v1 = r(Var)*((r(N)-1)/r(N))
summarize ips if zvar==0
scalar v0 = r(Var)*((r(N)-1)/r(N))
summarize zvar
scalar p1 = r(mean)
scalar p0 = 1-p1
scalar iv = (p1*v1*oe0+p0*v0*oe1)/(p1*v1*te0+p0*v0*te1)
```

Following Theorem 5, the outcome of this procedure (`iv`) will be identical to the coefficient on `tvar` in:

```
ivregress 2sls ovar xvars (tvar = zvar)
```

F Robustness Checks

Table 10: Alternative Estimates of the Effects of NSW Program

	(1)	(2)	(3)	(4)
	Matching on the LPM propensity score			
ATE	–9,227***	–7,504**	–6,245*	–6,581*
	(2,388)	(3,518)	(3,382)	(3,370)
ATT	–3,282***	257	975	–892
	(863)	(694)	(813)	(906)
ATC	–9,295***	–7,594**	–6,328*	–6,646*
	(2,415)	(3,556)	(3,420)	(3,409)
	Matching on the logit propensity score			
ATE	–6,682**	–7,683***	–4,187	–2,961
	(2,773)	(2,421)	(3,012)	(11,900)
ATT	–3,855***	265	2,117**	2,032**
	(854)	(695)	(856)	(860)
ATC	–6,714**	–7,775***	–4,260	–3,018
	(2,804)	(2,448)	(3,046)	(12,037)
	Oaxaca–Blinder			
ATE	–6,132***	–6,218**	–4,952*	–4,930
	(1,644)	(2,534)	(2,996)	(3,073)
ATT	–3,417***	–69	623	796
	(628)	(598)	(628)	(639)
ATC	–6,163***	–6,289**	–5,017*	–4,996
	(1,662)	(2,561)	(3,030)	(3,108)
Demographic controls	✓		✓	✓
"Earnings in 1974"				✓
Earnings in 1975		✓	✓	✓
Observations	16,177	16,177	16,177	16,177

Notes: See also LaLonde (1986), Dehejia and Wahba (1999), and Smith and Todd (2005) for more details on these data. The dependent variable is earnings in 1978. Demographic controls include age, age squared, years of schooling, and indicators for married, high school dropout, black, and Hispanic. For treated individuals, "Earnings in 1974" correspond to real earnings in months 13–24 prior to randomization, which overlaps with calendar year 1974 for a number of individuals. See Smith and Todd (2005) for further discussion. For "Matching on the LPM propensity score" and "Matching on the logit propensity score," estimation is based on nearest-neighbor matching on the estimated propensity score (with a single match). The propensity score is estimated using a linear probability model (LPM) or a logit model. For "Oaxaca–Blinder," estimation is based on the estimator discussed in Kline (2011). Huber–White standard errors (Oaxaca–Blinder) and Abadie–Imbens standard errors (matching) are in parentheses. Abadie–Imbens standard errors ignore that the propensity score is estimated.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table 11: Alternative Estimates of the Effects of Medieval Pogroms

	1920s pogroms	NSDAP 1928	DVFP 1924	Synagogue attacks
	Matching on the LPM propensity score			
ATE	0.0563**	0.0106*	0.0163	0.1079*
	(0.0225)	(0.0063)	(0.0128)	(0.0561)
ATT	0.0647***	0.0084	0.0100	0.1043*
	(0.0236)	(0.0061)	(0.0131)	(0.0584)
ATC	0.0341	0.0162*	0.0329*	0.1194*
	(0.0268)	(0.0099)	(0.0186)	(0.0706)
	Matching on the logit propensity score			
ATE	0.0531**	0.0095	0.0093	0.1295**
	(0.0222)	(0.0061)	(0.0130)	(0.0556)
ATT	0.0647***	0.0069	0.0093	0.1374**
	(0.0236)	(0.0062)	(0.0138)	(0.0573)
ATC	0.0227	0.0164*	0.0092	0.1045
	(0.0261)	(0.0087)	(0.0155)	(0.0725)
	Oaxaca–Blinder			
ATE	0.0578**	0.0128**	0.0131	0.1129**
	(0.0266)	(0.0052)	(0.0104)	(0.0469)
ATT	0.0555*	0.0117**	0.0118	0.1084**
	(0.0295)	(0.0051)	(0.0104)	(0.0457)
ATC	0.0639***	0.0155***	0.0164	0.1269**
	(0.0217)	(0.0060)	(0.0113)	(0.0544)
Log of population in 1924			✓	
Log of population in 1925	✓			
Log of population in 1928		✓		
Log of population in 1933				✓
% Jewish in 1925	✓	✓	✓	
% Jewish in 1933				✓
% Protestant in 1925	✓	✓	✓	✓
Observations	320	325	325	278

Notes: See also Voigtländer and Voth (2012) for more details on these data. The dependent variables are an indicator for pogroms during the 1920s, the vote share of the NSDAP in the May 1928 election, the vote share of the Deutsch-Völkische Freiheitspartei in the May 1924 election, and an indicator for whether a synagogue was destroyed or damaged in the 1938 *Reichskristallnacht*. For "Matching on the LPM propensity score" and "Matching on the logit propensity score," estimation is based on nearest-neighbor matching on the estimated propensity score (with a single match). The propensity score is estimated using a linear probability model (LPM) or a logit model. For "Oaxaca–Blinder," estimation is based on the estimator discussed in Kline (2011). Cluster-robust standard errors (Oaxaca–Blinder) and Abadie–Imbens standard errors (matching) are in parentheses. Abadie–Imbens standard errors ignore that the propensity score is estimated.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table 12: Alternative Estimates of the Effects of Catholic Schooling

	(1)	(2)	(3)	(4)
	Matching on the LPM propensity score			
ATE	1.31**	2.10***	—	—
	(0.56)	(0.51)	—	—
ATT	1.58***	1.44***	—	—
	(0.41)	(0.44)	—	—
ATC	1.29**	2.15***	—	—
	(0.58)	(0.53)	—	—
	Matching on the logit propensity score			
ATE	1.75***	2.10***	—	—
	(0.54)	(0.50)	—	—
ATT	1.48***	1.36***	—	—
	(0.40)	(0.43)	—	—
ATC	1.76***	2.15***	—	—
	(0.55)	(0.52)	—	—
	Oaxaca–Blinder			
ATE	1.80***	1.95***	—	—
	(0.43)	(0.44)	—	—
ATT	1.47***	1.59***	—	—
	(0.39)	(0.39)	—	—
ATC	1.82***	1.98***	—	—
	(0.44)	(0.45)	—	—
	Heckman’s two-step procedure			
ATE	—	—	–0.95	0.17
	—	—	(1.62)	(1.62)
ATT	—	—	3.99***	4.90***
	—	—	(1.39)	(1.35)
ATC	—	—	–1.27	–0.13
	—	—	(1.72)	(1.72)
Baseline controls	✓	✓	✓	✓
Demographic controls		✓		✓
Observations	7,444	7,306	7,444	7,306

Notes: See also Altonji *et al.* (2005) for more details on these data. The dependent variable is a standardized twelfth grade math test score. Baseline controls include mother’s education, father’s education, and log of family income. Demographic controls include indicators for female, Asian, black, Hispanic, a married parent, and a single-mother household. In columns 3 and 4, instrumental variables for Catholic schooling include the following indicators for distance from the nearest Catholic high school: less than 1 mile, between 1 and 3 miles, between 3 and 6 miles, and between 6 and 10 miles. For “Matching on the LPM propensity score” and “Matching on the logit propensity score,” estimation is based on nearest-neighbor matching on the estimated propensity score (with a single match). The propensity score is estimated using a linear probability model (LPM) or a logit model. For “Oaxaca–Blinder,” estimation is based on the estimator discussed in Kline (2011). For “Heckman’s two-step procedure,” estimation is based on the estimator discussed in Wooldridge (2015). Huber–White standard errors (Oaxaca–Blinder), Abadie–Imbens standard errors (matching), and bootstrap standard errors (Heckman’s two-step procedure) are in parentheses. Abadie–Imbens standard errors ignore that the propensity score is estimated.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table 13: Alternative Estimates of the Effects of Cash Transfers

	(1)	(2)	(3)	(4)
	Matching on the LPM propensity score			
ATE	0.0110	0.0147*	0.0022	0.0011
	(0.0070)	(0.0089)	(0.0099)	(0.0098)
ATT	0.0106	0.0143	–0.0002	–0.0002
	(0.0073)	(0.0096)	(0.0109)	(0.0107)
ATC	0.0144**	0.0179**	0.0194**	0.0100
	(0.0059)	(0.0082)	(0.0084)	(0.0085)
	Matching on the logit propensity score			
ATE	0.0111	0.0183**	–0.0019	–0.0054
	(0.0073)	(0.0081)	(0.0166)	(0.0166)
ATT	0.0107	0.0181**	–0.0043	–0.0105
	(0.0077)	(0.0087)	(0.0187)	(0.0186)
ATC	0.0145**	0.0193**	0.0152*	0.0309***
	(0.0059)	(0.0083)	(0.0085)	(0.0083)
	Oaxaca–Blinder			
ATE	0.0105*	0.0100	0.0140	0.0130
	(0.0063)	(0.0065)	(0.0100)	(0.0105)
ATT	0.0096	0.0092	0.0133	0.0124
	(0.0064)	(0.0068)	(0.0109)	(0.0115)
ATC	0.0164**	0.0160***	0.0184***	0.0170**
	(0.0069)	(0.0061)	(0.0071)	(0.0073)
State fixed effects	✓			
County fixed effects			✓	✓
Cohort fixed effects	✓	✓	✓	✓
State characteristics		✓	✓	✓
County characteristics		✓		
Individual characteristics		✓	✓	✓
Observations	7,860	7,859	7,859	7,857

Notes: See also Aizer *et al.* (2016) for more details on these data. The dependent variable is log age at death, as reported in the MP records (columns 1 to 3) or on the death certificate (column 4). State characteristics include manufacturing wages, age of school entry, minimum age for work permit, an indicator for a continuation school requirement, state laws concerning MP transfers (work requirement, reapplication requirement, and maximum amounts for first and second child), and log expenditures on education, charity, and social programs. County characteristics include average value of farm land, mean and SD of socio-economic index, poverty rate, female lfp rate, and shares of urban population, widowed women, children living with single mothers, and children working. Individual characteristics include child age at application, age of oldest and youngest child in family, number of letters in name, and indicators for the number of siblings, the marital status of the mother, and whether date of birth is incomplete. For "Matching on the LPM propensity score" and "Matching on the logit propensity score," estimation is based on nearest-neighbor matching on the estimated propensity score (with a single match). The propensity score is estimated using a linear probability model (LPM) or a logit model. For "Oaxaca–Blinder," estimation is based on the estimator discussed in Kline (2011). Cluster-robust standard errors (Oaxaca–Blinder) and Abadie–Imbens standard errors (matching) are in parentheses. Abadie–Imbens standard errors ignore that the propensity score is estimated.

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.