New Evidence on Linear Regression and Treatment Effect Heterogeneity

Słoczyński, Tymon

November 2014

New Evidence on Linear Regression and Treatment Effect Heterogeneity*

Tymon Słoczyński†

Abstract

It is standard practice in applied work to rely on linear least squares regression to estimate the effect of a binary variable ("treatment") on some outcome of interest. In this paper I study the interpretation of the regression estimand when treatment effects are in fact heterogeneous. I show that the coefficient on treatment is identical to the outcome of the following three-step procedure: first, calculate the linear projection of treatment on the vector of other covariates ("propensity score"); second, calculate average partial effects for both groups of interest from a regression of outcome on treatment, the propensity score, and their interaction; third, 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. Each of these steps is potentially problematic, but this last property—the reliance on implicit weights which are inversely related to the proportion of each group—can have particularly devastating consequences for applied work. To illustrate the severity of this issue, I perform Monte Carlo simulations as well as replicate two prominent applied papers: Berger, Easterly, Nunn and Satyanath (2013) on the effects of successful CIA interventions during the Cold War on imports from the US; and Martinez-Bravo (2014) on the effects of appointed officials on village-level electoral results in Indonesia. In both cases some of the conclusions change dramatically after allowing for heterogeneity in effects.

*This paper has benefited from many comments and discussions with Jeffrey Wooldridge, for which I am deeply grateful. I also thank Joshua Angrist, Marco Caliendo, Todd Elder, Macartan Humphreys, Guido Imbens, Krzysztof Karbownik, Nicholas Longford, Łukasz Marć, Michał Myck, Michela Tincani, Joanna Tyrowicz, Rudolf Winter-Ebmer, and seminar and conference participants at Ce(2) Workshop, CEPS/INSTEAD, IAAE Annual Conference, MEA and MEG Annual Meetings, Michigan State University, University of Potsdam, Upjohn Institute, Warsaw International Economic Meeting, Warsaw School of Economics, and WZB Berlin Social Science Center for useful comments and discussions. I acknowledge financial support from the National Science Centre (grant DEC-2012/05/N/HS4/00395), the Foundation for Polish Science (a START scholarship), and the "Weź stypendium—dla rozwoju" scholarship program.

†Warsaw School of Economics and IZA. E-mail: tymon.sloczynski@gmail.com.

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 (2014), 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

$$y_i = \alpha + \tau d_i + X_i\beta + \varepsilon_i, \tag{1}$$

where y denotes the outcome, d denotes the binary variable of interest, and X denotes the row vector of other covariates (control variables); $\hat{\tau}$ is then usually interpreted as the average treatment effect (ATE). This simple estimation strategy is used in recent papers by Fryer and Levitt (2004), Gittleman and Wolff (2004), Almond, Chay and Lee (2005), Elder, Goddeeris and Haider (2010), Fryer and Greenstone (2010), Fryer and Levitt (2010), Lang and Manove (2011), Alesina, Giuliano and Nunn (2013), Berger, Easterly, Nunn and Satyanath (2013), Bond and Lang (2013), Boustan and Collins (2013), Rothstein and Wozny (2013), Vogl (2013), Martinez-Bravo (2014), and many others.

The great appeal of linear least squares regression comes from its simplicity. 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 the model in (1). In this paper, therefore, I study the interpretation of the least squares estimand in the homogeneous linear model when treatment effects are in fact heterogeneous. I derive a new theoretical result which demonstrates that $\hat{\tau}$ is identical to the outcome of the following three-step procedure: in the first step, calculate the linear projection of d on X, *i.e.* the "propensity score" from the linear probability model; in the second step, regress 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. In consequence, when the proportion of one group *increases*, the weight on the effect on this group *decreases*. The limit of the regression estimand, as the proportion of treated units approaches unity, is the average treatment effect on the controls. I also establish conditions under which linear regression recovers

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

instead of

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

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

What follows, this paper contributes to a growing field of research in econometrics which studies the interpretation of various estimation methods when their underlying assumption of effect homogeneity is violated. See, for example, Wooldridge (2005), Løken, Mogstad and Wiswall (2012), Chernozhukov, Fernández-Val, Hahn and Newey (2013), Imai and Kim (2013), and Gibbons, Suárez Serrato and Urbancic (2014) for studies of fixed effects (FE) methods as well as Imbens and Angrist (1994), Angrist, Graddy and Imbens (2000), Løken *et al.* (2012), Kolesár (2013), and Dieterle and Snell (2014) for studies of instrumental variables (IV) estimators.[1] Also, the interpretation of the coefficient on a binary variable in linear least squares regression is studied by Angrist (1998) and Humphreys (2009), and both of these papers consider a saturated model for covariates, *i.e.* the estimating equation includes a separate binary variable for each combination of covariate values ("stratum").[2] In such a restricted setting, Angrist (1998) demonstrates that the weights underlying linear regression are proportional to the variance of treatment in each stratum.[3] Humphreys (2009) extends this result and shows that the linear regression estimand is bounded by both group-specific average treatment effects whenever treatment assignment probabilities are monotonic in stratum-specific effects. While both of these papers make substantive contributions, they might not always provide an accurate interpretation for linear regression estimates in applied studies, because saturated models are rarely used in practice.[4] In this paper, therefore, I complement these previous results by relaxing the saturated model restriction and still deriving a closed-form expression for

[1] 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 estimators, *conditional on X*, as weighted averages of marginal treatment effects.

[2] Also, the interpretation of the coefficient on a continuous variable in linear regression is studied by Yitzhaki (1996), Deaton (1997), Angrist and Krueger (1999), Løken *et al.* (2012), and Solon, Haider and Wooldridge (2013).

[3] A similar result for nonsaturated models is derived by Rhodes (2010) and Aronow and Samii (2014). In both of these papers the regression estimand is interpreted as a weighted average of individual-level treatment effects—which is quite different from this paper.

[4] For notable exceptions, see Angrist (1998), Black, Smith, Berger and Noel (2003), and Angrist and Pischke (2009).

the regression estimand—in terms of group-specific average treatment effects (τ_{ATT} and τ_{ATC}). This formulation is very attractive because each regression estimate can now be expressed as a weighted average of two estimates of τ_{ATT} and τ_{ATC}. Moreover, the weights are also easily computed—and they are always nonnegative and sum to one.

To illustrate the importance of this result, I perform Monte Carlo simulations and replicate two influential applied papers: Berger *et al.* (2013) and Martinez-Bravo (2014). Both of these papers study the effect of a binary variable (US interventions in foreign countries and whether the local officials are appointed or elected, respectively) on the expected value of some outcome of interest, and both rely on a model with homogeneous effects which is estimated using OLS. Berger *et al.* (2013) conclude that CIA interventions during the Cold War led to a dramatic increase in imports from the US, without affecting exports to the US, aggregate imports, and aggregate exports. However, when I present the implied estimates of the average effect of CIA interventions on intervened countries and nonintervened countries, it becomes clear that this conclusion is driven by the large discrepancy in the effect on nonintervened countries across specifications—while this parameter is arguably of little interest in this application.[5] The implied estimates of the average effect on intervened countries are all significantly positive and remarkably stable across specifications—and suggest that CIA interventions led to an (unbelievably large) increase in all measures of international trade in intervened countries. Surprisingly, when I relax the linear relationship between potential outcomes and the propensity score, and use a matching estimator, these effects often become significantly negative.

My second empirical application concentrates on the effects of appointed village heads on electoral results. In a recent paper, Martinez-Bravo (2014) studies the outcome of the first democratic election in Indonesia after the fall of the regime of General Soeharto. She concludes that Golkar, *i.e.* Soeharto's party, was more likely to win in *kelurahan* villages which had appointed village heads, compared with *desa* villages which had elected village heads. In this paper, however, I document that linear regression provides a very poor approximation to the average effect of appointed officials. Note that *kelurahan* villages constitute a small fraction of this data set, while my theoretical result suggests that linear regression will therefore attach nearly all of the weight to the average effect of appointed officials in these villages, and not in *desa*. This is reconfirmed in my analysis, and I conclude that the average treatment effect, *i.e.* the average difference in electoral results between similar *kelurahan* and *desa* villages, is not significantly different from zero.

[5]Imagine, for example, estimating the effect of CIA interventions in Australia, Canada, and the UK on their imports from the US. Note that the measure of CIA interventions equals one "if the CIA either installed a foreign leader or provided covert support for the regime once in power" (Berger *et al.*, 2013).

2 Theoretical Results

As before, let y denote the outcome, let d denote the binary variable of interest ("treatment"), and let X denote the row vector of other covariates. If $\mathrm{L}(\cdot \mid \cdot)$ denotes the linear projection, this paper is concerned with the interpretation of τ in

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

when the population linear model is possibly incorrect. Before giving my main theoretical results, however, I introduce further definitions. In particular, let

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

denote the unconditional probability of "treatment" and let

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

denote the "propensity score" from the linear probability model.[6] Note that $p(X)$ is the best linear approximation to the true propensity score. It is also helpful to introduce two linear projections of y on 1 and $p(X)$, separately for $d = 1$ and $d = 0$, namely

$$\mathrm{L}[y \mid 1, p(X)] = \alpha_1 + \gamma_1 \cdot p(X) \quad \text{if} \quad d = 1 \tag{7}$$

and also

$$\mathrm{L}[y \mid 1, p(X)] = \alpha_0 + \gamma_0 \cdot p(X) \quad \text{if} \quad d = 0. \tag{8}$$

Note that Equations 6–8 are definitional. I do not assume that these linear projections correspond to well-specified population models and I do not put any restrictions on the underlying data-generating process. Similarly, I define the average partial effect of d as

$$\tau_{APE} = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot \mathrm{E}[p(X)] \tag{9}$$

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

$$\tau_{APE|d=j} = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot \mathrm{E}[p(X) \mid d = j]. \tag{10}$$

[6]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.

If d is unconfounded conditional on X, then the propensity score theorem (Rosenbaum and Rubin, 1983) implies that τ_{APE}, $\tau_{APE|d=1}$, and $\tau_{APE|d=0}$ have a useful interpretation as the average treatment effect, the average treatment effect on the treated, and the average treatment effect on the controls, respectively. It should be stressed, however, that the main result of this paper (Theorem 1) is more general and does not require unconfoundedness.

Theorem 1 (Decomposition of the Linear Regression Estimand) *Define τ as in (4) and define $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$ as in (10). Let $\mathrm{V}(\cdot \mid \cdot)$ denote the conditional variance. Then,*

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

Theorem 1 shows that τ, the linear regression estimand, can be expressed as a weighted average of $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$, with nonnegative weights which always sum to one.[7] The definition of $\tau_{APE|d=j}$ makes it clear that the regression estimand is always 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". In applied work, however, it is quite rare to estimate propensity scores using the linear probability model, probably because the estimated probabilities are not ensured to be strictly between zero and one—and therefore it is important to note that linear regression is implicitly based on this procedure. Next, in the second step, we obtain $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$ from a regression of y on d, $p\left(X\right)$, and their interaction. Again, similar procedures are rarely used in practice and are generally not recommended, because it is difficult to motivate a linear relationship between potential outcomes and the propensity score (see, *e.g.*, Imbens and Wooldridge, 2009). According to Theorem 1, however, linear regression is implicitly based on this restrictive model. 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 linear regression on $\tau_{APE|d=1}$ is increasing in $\mathrm{V}\left[p\left(X\right) \mid d=0\right]$ and $1-\rho$ and the weight which is placed on $\tau_{APE|d=0}$ is increasing in $\mathrm{V}\left[p\left(X\right) \mid d=1\right]$ and ρ.

At first, this weighting scheme might be seen as surprising: the more units belong to group j ($d=j$, $j=0,1$), the less weight is placed on $\tau_{APE|d=j}$, *i.e.* the effect *on this group*. To aid intuition, recall that the linear regression model is based on the assumption of homogeneity in effects; in particular, $\tau_{APE} = \tau_{APE|d=1} = \tau_{APE|d=0}$. Notice also that $\tau_{APE|d=1}$ ($\tau_{APE|d=0}$) is estimated, in general, using the data from units with $d=0$ ($d=1$).

[7] See Appendix A for the proof of Theorem 1.

Therefore, if effects are assumed to be homogeneous, we want to place more (less) weight on $\hat{\tau}_{APE|d=1}$ when the proportion of units with $d = 1$ decreases (increases), as this will improve efficiency in estimating τ_{APE}. However, the opposite holds true if effects are allowed to be heterogeneous, and then using linear regression is likely to introduce bias.

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

Corollary 1 (Causal Interpretation of the Linear Regression Estimand) *Suppose that d is unconfounded conditional on X and that the population models for d and y are linear in X and $p(X)$, respectively. Then, Theorem 1 implies that*

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

In other words, if one assumes that unconfoundedness holds and that the population models for d and y are correctly specified as linear in X and $p(X)$, respectively, the weighting scheme from Theorem 1 will apply to τ_{ATT} and τ_{ATC}. In particular, the weight which is placed on τ_{ATT} is increasing in $1-\rho$ and the weight which is placed on τ_{ATC} is increasing in ρ. Corollary 2 shows that the relationship between τ and ρ is in fact monotonic. The only case where τ is unrelated to ρ occurs when both group-specific average partial effects are equal.

Corollary 2 *Theorem 1 implies that*

$$\frac{\mathrm{d}\tau}{\mathrm{d}\rho} = \frac{\mathrm{V}\left[p(X) \mid d=1\right] \cdot \mathrm{V}\left[p(X) \mid d=0\right] \cdot \left[\tau_{APE|d=0} - \tau_{APE|d=1}\right]}{\left[\rho \cdot \mathrm{V}\left[p(X) \mid d=1\right] + (1-\rho) \cdot \mathrm{V}\left[p(X) \mid d=0\right]\right]^2}.$$

Therefore, if $\tau_{APE|d=1} > \tau_{APE|d=0}$, then $\frac{\mathrm{d}\tau}{\mathrm{d}\rho} < 0$. With an increase in ρ, τ deviates from $\tau_{APE|d=1}$ towards $\tau_{APE|d=0}$. Similarly, if $\tau_{APE|d=1} < \tau_{APE|d=0}$, then $\frac{\mathrm{d}\tau}{\mathrm{d}\rho} > 0$. Again, with an increase in ρ, τ deviates from $\tau_{APE|d=1}$ towards $\tau_{APE|d=0}$. In other words, when $\tau_{APE|d=1} \neq \tau_{APE|d=0}$ and the proportion of one group changes, the weight on the effect on this group always changes in the opposite direction.

Corollary 3 *Theorem 1 implies that*

$$\lim_{\rho \to 1} \tau = \tau_{APE|d=0} \qquad \textit{and} \qquad \lim_{\rho \to 0} \tau = \tau_{APE|d=1}.$$

According to Corollary 3, another consequence of Theorem 1 is that the linear regression estimand approaches the average partial effect on group j whenever—in the limit—the proportion of units with $d = j$ goes to zero. Under a causal interpretation, when nearly everyone is treated, we get very close to the average treatment effect on the controls; conversely, when nearly nobody gets treated, we approach the average treatment effect on this (nearly nonexistent) group. Therefore, Corollary 3 provides the foundation for a simple rule of thumb: if nearly everyone belongs to group j, linear regression will approximately provide an estimate of the effect *on the other group*. As noted previously, this is a reasonable property under the assumption of homogeneity in effects: if nearly everyone belongs to group j, then we can estimate the effect on the other group, and not on group j, with relative precision. This argument arises from the fact that we use the data from units with $d = 0$ ($d = 1$) to estimate the counterfactual for units with $d = 1$ ($d = 0$); therefore, the precision of the estimates for group j is increasing in the amount of data from the other group. If we maintain the assumption of homogeneity in effects, then we should indeed place little weight on the effect for the large group. This logic, however, is no longer applicable when effects are allowed to be heterogeneous.

Another consequence of Theorem 1 is described by Corollary 4. We can start with noting that the average partial effect of d can be written as

$$\tau_{APE} = \rho \cdot \tau_{APE|d=1} + (1 - \rho) \cdot \tau_{APE|d=0}. \tag{11}$$

Then, Corollary 4 provides a condition under which linear regression reverses these "natural" weights on $\tau_{APE|d=1}$ and $\tau_{APE|d=0}$.

Corollary 4 *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 linear regression estimand is equal to a weighted average of both group-specific average 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 4 allows the linear regression estimand to be equal to the average partial effect of d, and this occurs 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 $\rho = 1 - \rho = \frac{1}{2}$. Moreover, Corollary 5 provides a more general condition under which we can recover the average partial effect of d using linear regression.

Corollary 5 *Suppose that $\tau_{APE|d=1} \neq \tau_{APE|d=0}$. Then, Theorem 1 implies that*

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

Of course, this condition is very demanding, and we cannot, in general, expect it to hold. Corollary 5 can therefore be seen as an alternative example of the "knife-edge special case" of consistency of OLS, similar to Solon *et al.* (2013).

So how can we solve the problem described in Theorem 1? Actually, there are many well-known estimation methods which do not pose similar problems. First, it is sufficient to interact the binary variable of interest with other covariates, and then calculate the average partial effect of d on a particular group (similar to Equations 9 and 10). 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 simply "regression" (Imbens and Wooldridge, 2009). Second, one can 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 (WLS). In particular, we might use the method of Lin (2013), in which Equation 1 is estimated using WLS, with weights of $\frac{1-\rho}{\rho}$ for units with $d = 1$ and weights of $\frac{\rho}{1-\rho}$ for units with $d = 0$. However, note that—unlike in Lin (2013) who studies regression adjustments to experimental data—this estimator is consistent for the average partial effect of d only in a special case, namely under the restrictive condition in Corollary 4, $\mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right]$, which is trivially true in an experimental setting, but not in a nonexperimental study.[8]

[8]The crucial difference between regression adjustment in a setting with experimental data and in a setting with nonexperimental data comes from the fact that—under a causal interpretation—the average treatment effect on the treated and the average treatment effect on the controls are necessarily equal—in expectation—in a randomized experiment, but not in a nonexperimental study. See Freedman (2008a,b),

3 Monte Carlo

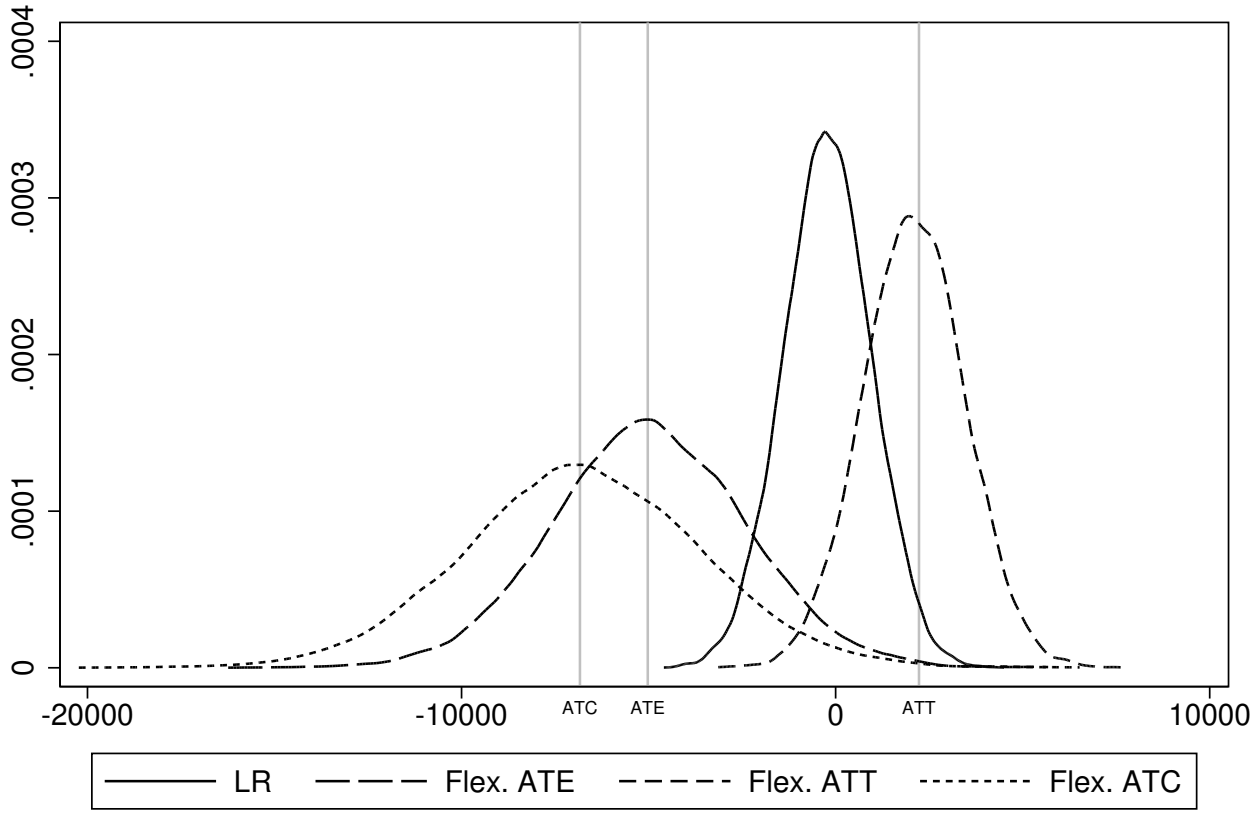
This section illustrates some of the key ideas of this paper using two Monte Carlo studies. The first study is similar to that in a recent paper by Busso, DiNardo and McCrary (2013), and it also attempts to mimic some features of the data from the National Supported Work (NSW) Demonstration (LaLonde, 1986). As in Busso *et al.* (2013), I focus on the subsample of African Americans as well as the comparison sample from the Panel Study of Income Dynamics (PSID), also restricted to African Americans. The outcome of interest is earnings in 1978, and the vector of covariates includes age, years of education, an indicator for being a high school dropout, marital status, earnings in 1974, earnings in 1975, employment status in 1974, and employment status in 1975. There are 156 treated units and 624 control units in the final data set. In the first step, I estimate a probit model for treatment, and calculate a linear prediction from this model ("propensity score"). For each treatment status, I also estimate a regression model for outcome, and again calculate predicted values. In the second step, I draw with replacement 780 vectors which consist of: a vector of covariates, predicted values of both potential outcomes, and the estimated propensity score. In the third step, I draw iid normal errors, and use them—together with the estimated propensity score—to construct a treatment status for each unit. In the fourth step, separately for each treatment status, I draw iid normal errors, and use them—together with predicted values from both regression models—to construct potential outcomes for each unit. Finally, for each unit, the treatment status is used to determine which potential outcome is observed.

This procedure is used to draw 10,000 hypothetical samples. For each sample, I estimate the effect of treatment using linear least squares regression—and then calculate the estimates of the average treatment effect on the treated and the average treatment effect on the controls which are implied by Theorem 1. I also calculate the implicit weights on these estimates. Moreover, I estimate the average treatment effect, the average treatment effect on the treated, and the average treatment effect on the controls using the "flexible OLS" estimator—which is expected to be unbiased, given the data-generating process described above. It might also be useful to note that the true values of these parameters are equal to –$5,022, $2,229, and –$6,835, respectively.

The main results of this Monte Carlo study are summarized in Figure 1. Each of the "flexible OLS" estimators is unbiased for its respective parameter. At the same time, however, linear regression is very biased for each of τ_{ATE}, τ_{ATT}, and τ_{ATC}, with the smallest

Deaton (2010), Schochet (2010), and Lin (2013) for recent discussions of regression adjustments to experimental data.

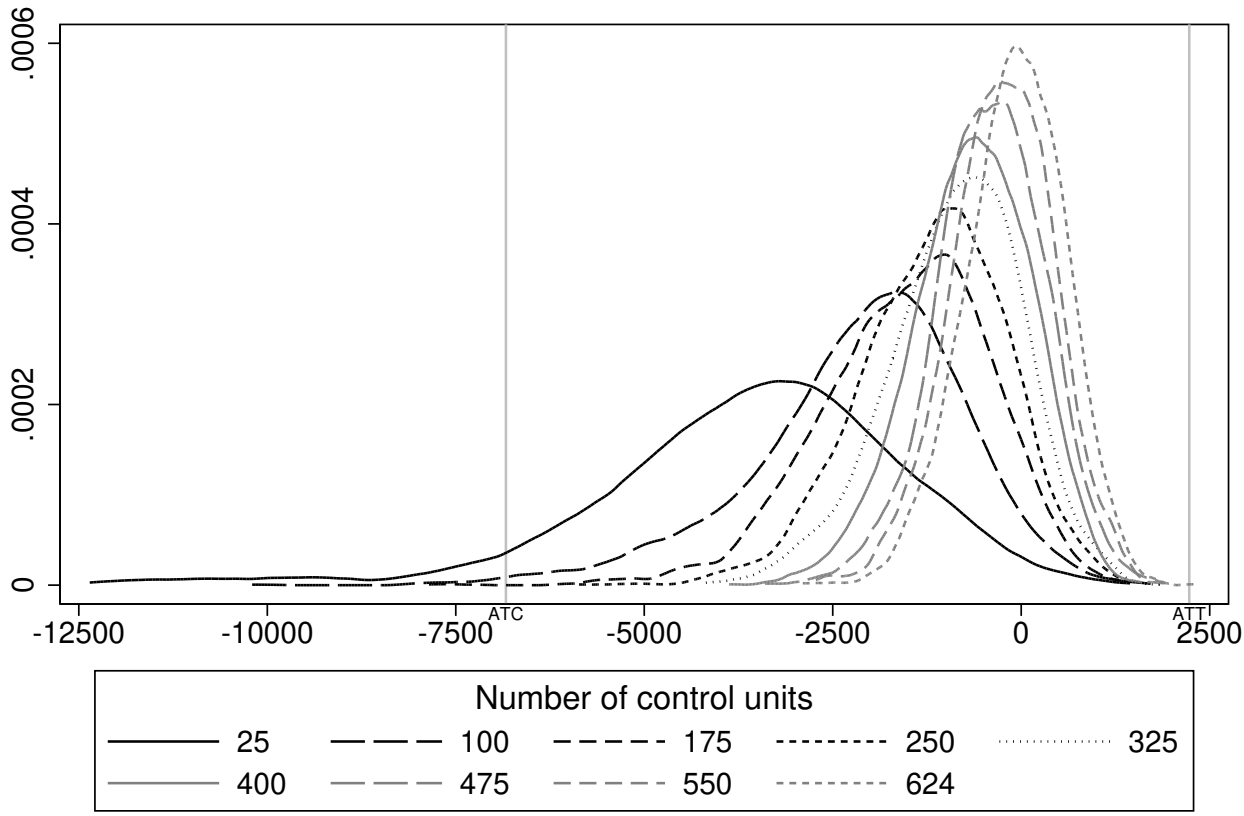
Figure 1: Linear Regression and "Flexible OLS" Estimates of Average Treatment Effects



bias in estimating τ_{ATT} (for more details, see Table B7 in Appendix B). Note that, on average, only 20% of the units are treated. Consequently, linear regression is usually closest to the true effect on the treated, *the smaller group*, although it is still biased for this parameter. Given Theorem 1, this result should not seem surprising.

Additional results are presented in Appendix B. In particular, Figure B3 and Table B7 provide evidence of poor finite-sample performance of both components of linear regression, *i.e.* the LPM-based estimators of the average treatment effect on the treated and the average treatment effect on the controls, which are implied by Theorem 1. It is clear that both of these estimators—unlike the "flexible OLS" estimators in Figure 1—are biased for their respective parameters, given the data-generating process in this Monte Carlo study. This is most easily visible in Figure B3. Moreover, Table B8 summarizes the empirical distribution of the implicit weights which are used by linear regression to reweight both of these estimates. Even though, on average, 20% of the units are treated, the average weight on $\hat{\tau}_{ATT}$ is 0.640, with the standard deviation of 0.038 (across 10,000 replications). In other words, under partial effect heterogeneity linear least squares regression is equivalent to a

Figure 2: Linear Regression Estimates for Different Values of N_0



weighted average of two estimators, both of which are likely to perform poorly in finite samples, with weights which are also poorly chosen. It would be difficult to motivate the use of linear least squares regression under similar circumstances.

The second simulation study is based on the same sample of African Americans, and it uses a variant of the nonparametric bootstrap. In each replication, I retain the original sample of 156 treated units. I also draw a subsample of size N_0, with replacement, from the original sample of 624 control units—and append it to the sample of treated units. Importantly, I consider nine values of N_0: 25, 100, 175, 250, 325, 400, 475, 550, and 624. For each N_0, I draw 2,500 hypothetical samples, and then examine the effects of N_0 on the finite-sample performance of linear least squares regression.

The results are summarized in Figure 2.[9] An obvious conclusion is that the higher the proportion of control units, the further we get from the average treatment effect on the controls—and closer to the average treatment effect on the treated. This relationship is monotonic, as previously noted in Corollary 2. Additional results from this simulation

[9]For clarity, Figure 2 excludes 32 estimates (less than 0.15% of their total number) which are smaller than –12,500. Of course, this exclusion has no effect on the interpretation of this figure.

study are presented, again, in Appendix B. In particular, Table B9 shows the mean and median bias, the root-mean-square error (RMSE), the median-absolute error (MAE), and the standard deviation of linear least squares regression—separately for each N_0 and for the estimation of τ_{ATT} and τ_{ATC}. The conclusions are, of course, the same: in terms of bias, RMSE, and MAE, the performance of linear regression in estimating τ_{ATT} improves with the proportion of control units; similarly, when the proportion of treated units increases, we get closer to τ_{ATC}. Moreover, Table B10 summarizes the empirical distribution of the implicit weights which are used by linear least squares regression to reweight the implied estimates of τ_{ATT} and τ_{ATC}—again, separately for each N_0. When the proportion of treated units varies between 0.200 and 0.862, the average weight on $\hat{\tau}_{ATT}$ varies between 0.638 and 0.368; it is therefore useful to note that—at least in this particular simulation study—the average weights on $\hat{\tau}_{ATT}$ and $\hat{\tau}_{ATC}$ vary somewhat less than the proportions of both groups, but there is also significant variation in weights for each value of N_0. However, as evident in Table B10, the negative relationship between the proportion of treated (control) units and the implicit weight on $\hat{\tau}_{ATT}$ ($\hat{\tau}_{ATC}$) is generally very strong.

4 Empirical Applications

This section illustrates the importance of the main theoretical result of this paper by means of a replication of two applied papers: Berger *et al.* (2013) on the effects of CIA interventions during the Cold War on imports from the US; and Martinez-Bravo (2014) on the effects of appointed officials on village-level electoral results in Indonesia.

The Effects of US Influence on International Trade (Berger *et al.*, 2013)

In a recent paper, Berger *et al.* (2013) provide evidence that successful CIA interventions during the Cold War were used to create a larger foreign market for US-produced goods. The authors use recently declassified CIA documents to construct country- and year-specific measures of US political influence, and conclude that such influence had a positive effect on the share of total imports that intervened countries purchased from the US. At the same time, however, Berger *et al.* (2013) find no evidence that CIA interventions increased exports to the US, total imports, or total exports.

In this study, the treatment variable ("CIA intervention" or "US influence") is binary, and equals one whenever—in a given country and year—the CIA either installed a new leader or provided support for the current regime. These activities took various forms, and included "creation and dissemination of (often false) propaganda, ...covert politi-

Table 1: A Replication of Berger *et al.* (2013)

	ln imports (US)	ln imports (US)	ln imports (US)	ln imports (world)	ln exports (US)	ln exports (world)
CIA intervention	0.283**	0.776***	0.293***	–0.009	0.058	0.000
	(0.110)	(0.143)	(0.109)	(0.045)	(0.122)	(0.052)
Country fixed effects	✓		✓	✓	✓	✓
Trade costs and MR controls		✓	✓	✓	✓	✓
Observations	4,149	4,149	4,149	4,149	3,922	3,922

Notes: See also Berger *et al.* (2013) for more details on these data. The unit of observation is a country c in year t, where c excludes the US and the Soviet Union and t ranges between 1947 and 1989. The dependent variables are listed in the column headings. Exact definitions of these variables are given in Berger *et al.* (2013). All regressions include year fixed effects, a Soviet intervention control, ln per capita income, an indicator for leader turnover, current leader tenure, and a democracy indicator. Estimation is based on linear least squares regression. Newey–West standard errors are in parentheses.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

cal operations, . . . the destruction of physical infrastructure and capital, as well as covert paramilitary operations" (Berger *et al.*, 2013). Apart from the treatment variable, the authors also control for year fixed effects, a Soviet intervention control, ln per capita income, an indicator for leader turnover, current leader tenure, as well as a democracy indicator. The majority of their baseline specifications also include country fixed effects, trade costs, and Baier–Bergstrand multilateral resistance (MR) terms. The final sample consists of 166 countries, excludes the US and the Soviet Union, and covers the period from 1947 to 1989. Among the 166 countries, 51 experienced a CIA intervention during the Cold War. In a typical year, successful CIA interventions were taking place in 25 countries.

Table 1 reproduces the baseline estimates from Berger *et al.* (2013). Columns 1–3 report the estimated effects of CIA interventions on imports from the US. All of the coefficients are positive and statistically significant. The estimates from columns 1 and 3 are also very similar in magnitude; the estimate from column 2 is much larger, but this specification excludes country fixed effects. Therefore, Berger *et al.* (2013) conclude that CIA interventions increased US imports by almost 30 log points (as in columns 1 and 3), and then their remaining specifications control for country fixed effects, trade costs, and MR controls. Further estimates—for different dependent variables—are reported in columns 4–6. All of these coefficients are insignificant and very close to zero. The authors conclude that CIA interventions had no impact on exports to the US, total imports, or total exports.

Perhaps surprisingly, however, the authors interpret their main coefficient of interest as "the average reduced-form impact of CIA interventions on the countries that experience an intervention" (Berger *et al.*, 2013). Unfortunately, this is not a correct interpretation, given their reliance on a model with homogeneous effects which is estimated

Table 2: Berger *et al.* (2013) and Treatment Effect Heterogeneity

	ln imports (US)	ln imports (US)	ln imports (US)	ln imports (world)	ln exports (US)	ln exports (world)
CIA intervention	0.283**	0.776***	0.293***	–0.009	0.058	0.000
	(0.110)	(0.143)	(0.109)	(0.045)	(0.122)	(0.052)
Decomposition (Theorem 1)						
a. ATT	0.648***	0.794***	0.717***	0.691***	0.665***	0.863***
	(0.138)	(0.059)	(0.142)	(0.150)	(0.169)	(0.145)
b. w_{ATT}	0.676	0.832	0.677	0.678	0.689	0.691
c. ATC	–0.478***	0.691***	–0.595***	–1.484***	–1.288***	–1.928***
	(0.144)	(0.073)	(0.145)	(0.167)	(0.192)	(0.183)
d. w_{ATC}	0.324	0.168	0.323	0.322	0.311	0.309
OLS $= a \cdot b + c \cdot d$	0.283**	0.776***	0.293***	–0.009	0.058	0.000
	(0.110)	(0.143)	(0.109)	(0.045)	(0.122)	(0.052)
Country fixed effects	✓		✓	✓	✓	✓
Trade costs and MR controls		✓	✓	✓	✓	✓
Observations	4,149	4,149	4,149	4,149	3,922	3,922
P $(d = 1)$	0.225	0.225	0.225	0.225	0.235	0.235

Notes: See also Berger *et al.* (2013) for more details on these data. The unit of observation is a country c in year t, where c excludes the US and the Soviet Union and t ranges between 1947 and 1989. The dependent variables are listed in the column headings. Exact definitions of these variables are given in Berger *et al.* (2013). All regressions and propensity score specifications include year fixed effects, a Soviet intervention control, ln per capita income, an indicator for leader turnover, current leader tenure, and a democracy indicator. Estimation of "CIA intervention" (=OLS) is based on linear least squares regression. Estimation of ATT and ATC is described in Section 2 (in particular, see Theorem 1). Newey–West standard errors (OLS) and Huber–White standard errors (ATT and ATC) are in parentheses. Huber–White standard errors ignore that the propensity score is estimated.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

using ordinary least squares. An interpretation is given, however, in Theorem 1 in this paper: the estimates in Table 1 are all weighted averages of the average effect of CIA interventions on intervened countries (ATT) and the average effect of CIA interventions on nonintervened countries (ATC), with weights which are perhaps poorly chosen. At the same time, it is certainly very convincing to follow the intention of the authors, and focus on the average effect on intervened countries. This parameter can be used to answer the question about the actual consequences of CIA interventions during the Cold War. It is less useful to estimate the effect of CIA interventions on countries, in which interventions were highly unlikely, such as Australia, Canada, or the United Kingdom. Therefore, the average effect on nonintervened countries is arguably of little interest in this application, and I focus on the average effect on the "treated".

Table 2 decomposes the baseline estimates from Berger *et al.* (2013) into two components, the average effect of CIA interventions on intervened countries (ATT) and the av-

erage effect of CIA interventions on nonintervened countries (ATC). It also reports the implicit weights on these estimates. First, it is useful to note that about 23% of the units are treated, but at the same time the weight on $\hat{\tau}_{ATT}$ varies between 0.676 and 0.832. Second, the implied estimates of the average effect of CIA interventions on intervened countries are all positive, statistically significant, and very similar in magnitude. These estimates suggest that CIA interventions influenced all measures of international trade, and increased US imports, US exports, total imports, and total exports by 65–86 log points. Therefore, the large discrepancies in the estimates reported in Table 1—and the main conclusion in Berger *et al.* (2013)—are driven by the large variation in the effect on nonintervened countries across specifications. This is easily visible in Table 2, where $\hat{\tau}_{ATC}$ varies between –1.928 and 0.691, and hence we get the reported variation in the OLS estimate. Whenever $\hat{\tau}_{ATC}$ is negative and relatively large in absolute value (columns 4–6), the weighted average of $\hat{\tau}_{ATC}$ and $\hat{\tau}_{ATT}$ is approximately zero. Whenever $\hat{\tau}_{ATC}$ is relatively close to zero (columns 1–3), this weighted average becomes significantly positive.

Still, the following question arises: did CIA interventions really increase international trade in intervened countries by 65–86 log points? The magnitude of this effect is arguably difficult to believe, and we need to recall that these estimates are based on an estimator which is likely to perform very poorly in finite samples (for more details, see Section 3). More precisely, this method involves two steps: in the first step, calculate the "propensity score" from the linear probability model; in the second step, calculate average partial effects from a model which assumes a linear relationship between potential outcomes and this "propensity score". This second linearity assumption is particularly restrictive, and therefore we might need an additional robustness check.

Table 3 reports nearest-neighbor matching estimates of the average effect of CIA interventions on intervened countries. I consider two alternative models for the propensity score: a linear probability model and a probit model. In the first case, I use the same model as in the previous procedure, and retain the estimates of the "propensity score" from Table 2. In other words, I relax a restrictive assumption from the second stage of the previous two-step procedure, but retain the first stage. As evident in Table 3, the previous estimates of the average effect of CIA interventions on intervened countries are not robust to relaxing this assumption. The majority of the estimates become negative and often statistically significant.

In the second case, I use a probit model for the propensity score, but also implement an additional refinement of the matching procedure—namely, a requirement of exact matching within each country. As evident in Table 3, again, the estimates are not robust to this change in the procedure. Columns 1–3 report the estimated effects of CIA interven-

Table 3: Matching Estimates of the Effects of US Influence on International Trade

	ln imports (US)	ln imports (US)	ln imports (US)	ln imports (world)	ln exports (US)	ln exports (world)
ATT-LPM	–0.839	0.823***	–0.905*	–0.915*	–1.533**	–0.817
	(0.558)	(0.085)	(0.527)	(0.528)	(0.610)	(0.529)
Observations	4,149	4,149	4,149	4,149	3,922	3,922
ATT-probit	–0.130	–0.212	–0.193	–0.684***	–0.208	–0.572**
	(0.207)	(0.190)	(0.211)	(0.241)	(0.308)	(0.247)
Observations	1,406	1,406	1,406	1,406	1,318	1,318
Country fixed effects	✓		✓	✓	✓	✓
Trade costs and MR controls		✓	✓	✓	✓	✓

Notes: See also Berger *et al.* (2013) for more details on these data. The unit of observation is a country c in year t, where c excludes the US and the Soviet Union and t ranges between 1947 and 1989. The dependent variables are listed in the column headings. Exact definitions of these variables are given in Berger *et al.* (2013). All propensity score specifications include year fixed effects, a Soviet intervention control, ln per capita income, an indicator for leader turnover, current leader tenure, and a democracy indicator. Estimation is based on nearest-neighbor matching on the estimated propensity score (with a single match). For "ATT-probit", exact matching on c is also required. The propensity score is estimated using a linear probability model ("ATT-LPM") or a probit model ("ATT-probit"). Abadie–Imbens standard errors are in parentheses. These standard errors ignore that the propensity score is estimated.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

tions on US imports. All of these estimates are insignificant and close to zero. Similarly, the estimated effect on US exports is also small and statistically insignificant (column 5). However, the estimated effects on total imports and total exports (columns 4 and 6) are both negative, statistically significant, and similar in magnitude. These results lead to an alternative interpretation of these data: CIA interventions might have had a negative effect on international trade in intervened countries, perhaps by means of destabilizing their economies and their political institutions. At the same time, the estimated effects on US imports and US exports are much smaller than on total imports and total exports. Presumably, successful CIA interventions during the Cold War were indeed used to determine international trade in intervened countries—and counterbalance the negative effects of these interventions on US trade—but the pattern of these effects is likely to be very different from the interpretation in Berger *et al.* (2013).

The Effects of Local Officials on Electoral Results (Martinez-Bravo, 2014)

A recent paper by Martinez-Bravo (2014) examines—both theoretically and empirically—the differences in behavior between appointed and elected officials. In particular, the author focuses on the 1999 parliamentary election in Indonesia, *i.e.* on the first democratic election in this country after the fall of the Soeharto regime, and compares the electoral

results in *kelurahan* and in *desa* villages (which have appointed and elected heads, respectively). She concludes that Golkar, *i.e.* Soeharto's party, was significantly more likely to win in *kelurahan* than in *desa* villages, and hence that "the body of appointed officials ... is a key determinant of the extent of electoral fraud and clientelistic spending in new democracies" (Martinez-Bravo, 2014).

The treatment variable is again binary—and equals one for *kelurahan* villages. The sample consists of 43,394 villages, of which 3,036 (7%) are *kelurahan* and 40,358 (93%) are *desa*. The outcome variable is also binary, and equals one if Golkar was the most voted party in the village; in some cases—though not in the baseline specifications—there is an alternative outcome variable, which equals one if PDI-P (a competing party and the winner of the 1999 election) was the most voted party in this village. The majority of specifications also include district (*kabupaten*) fixed effects, and many specifications control for various geographical characteristics of the villages as well as for the availability of religious, health, and educational facilities.

It is important to note that Martinez-Bravo (2014) does not specify whether her intention is to estimate the average effect of appointed officials (ATE) or the average effect of appointed officials on *kelurahan* villages (ATT). Both of these parameters are potentially interesting, although the former is presumably more in line with one of the main objectives of Martinez-Bravo (2014), *i.e.* testing for (average) differences in behavior between appointed and elected officials. The latter parameter would be more relevant if our intention was to examine the actual impact of appointed officials on the electoral outcome. Therefore, in this section, I focus on the average treatment effect, but discuss various estimates of both this parameter and the average effect on the "treated".

Recall, however, that neither of these parameters is recovered by linear least squares regression, while this is the primary estimation method used by Martinez-Bravo (2014). The author also uses a probit model and a particular method based on the propensity score, and all these methods seem to give similar answers. However, quite unexpectedly, this particular propensity-score method—used by Martinez-Bravo (2014)—is implicitly based on the assumption of homogeneity in effects; it is in fact equivalent to a variant of linear least squares regression with a different set of control variables. More precisely, this method involves three steps: in the first step, the author estimates the propensity score using an algorithm based on a probit model; in the second step, she imposes the overlap condition, calculates quintiles of the distribution of the estimated propensity score, and uses them to generate five propensity-score strata; in the third step, she runs the regression of the dependent variable on the *kelurahan* indicator, province fixed effects, five indicator variables for the strata, and the full set of interactions between the strata and

Table 4: A Replication of Martinez-Bravo (2014)

	Linear probability model					Propensity score model		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Kelurahan indicator	0.074***	0.006	0.057***	0.057***	0.055***	0.023***	0.030***	0.033***
	(0.028)	(0.012)	(0.012)	(0.012)	(0.012)	(0.008)	(0.009)	(0.008)
Geographic controls			✓	✓	✓	✓	✓	✓
Religious controls				✓	✓		✓	✓
Facilities controls					✓			✓
District fixed effects		✓	✓	✓	✓	✓	✓	✓
Observations	43,394	43,394	43,394	43,394	43,394	21,502	20,565	19,206

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if Golkar was the most voted party in the village in the 1999 parliamentary election and zero otherwise. Geographic controls include population density, a quartic in the logarithm of the village population, a quartic in the percentage of households whose main occupation is in agriculture, share of agricultural land in the village, distance to the subdistrict office, distance to the district capital, and indicators for urban and high altitude. Religious controls include the number of mosques, prayer houses, churches, and Buddhist temples per 1,000 people. Facilities controls include the number of hospitals, maternity hospitals, polyclinics, *puskesmas* (primary care centers), kindergartens, primary schools, high schools, and TVs per 1,000 people. Estimation is based on linear least squares regression, with controls for either the variables listed in the table (columns 1–5) or the propensity-score strata, province fixed effects, and the full set of interactions between the strata and the fixed effects (columns 6–8). In the latter case, the variables listed in the table correspond to the propensity score specifications. Cluster-robust standard errors (columns 1–5) and bootstrap standard errors (columns 6–8) are in parentheses.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

the fixed effects. Because this last regression does not include interactions between the control variables and the treatment variable, Martinez-Bravo (2014) implicitly makes the assumption of treatment effect homogeneity.

Consequently, Table 4 reproduces the baseline estimates from Martinez-Bravo (2014), both for the linear probability model and for the propensity score model. There are large differences between the coefficients in column 1 and 2 as well as between column 2 and 3. However, when geographic controls are included in column 3, the estimated effect stabilizes, and suggests that appointed officials increased the probability of Golkar victory by 6 percentage points (columns 3–5) or 2–3 percentage points (columns 6–8). All of these coefficients are statistically significant and also very similar in magnitude within each of the estimation methods.

Table 5 applies the main theoretical result of this paper to these estimates, and decomposes all the baseline coefficients from Martinez-Bravo (2014) into two components, the average effect of appointed officials on *kelurahan* villages (ATT) and the average effect of appointed officials on *desa* villages (ATC). I also report the implicit weights which are used by linear regression to reweight both of these estimates. While the proportion of *kelurahan* villages varies between 7% and 12%, the weight on $\hat{\tau}_{ATT}$ varies between 0.490

Table 5: Martinez-Bravo (2014) and Treatment Effect Heterogeneity

	Linear probability model					Propensity score model		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Kelurahan indicator	0.074***	0.006	0.057***	0.057***	0.055***	0.023***	0.030***	0.033***
	(0.028)	(0.012)	(0.012)	(0.012)	(0.012)	(0.008)	(0.009)	(0.008)
Decomposition (Theorem 1)								
a. ATT		–0.064*	–0.008	–0.008	–0.009	0.037**	0.045***	0.045***
		(0.037)	(0.028)	(0.028)	(0.028)	(0.016)	(0.016)	(0.016)
b. w_{ATT}		0.490	0.671	0.672	0.679	0.785	0.788	0.779
c. ATC		0.074***	0.192***	0.191***	0.192***	–0.026	–0.029	–0.011
		(0.027)	(0.041)	(0.041)	(0.042)	(0.032)	(0.034)	(0.032)
d. w_{ATC}		0.510	0.329	0.328	0.321	0.215	0.212	0.221
OLS $= a \cdot b + c \cdot d$	0.074***	0.006	0.057***	0.057***	0.055***	0.023***	0.030***	0.033***
	(0.028)	(0.012)	(0.012)	(0.012)	(0.012)	(0.008)	(0.009)	(0.008)
e. P $(d = 1)$		0.070	0.070	0.070	0.070	0.112	0.114	0.116
f. P $(d = 0)$		0.930	0.930	0.930	0.930	0.888	0.886	0.884
ATE $= e \cdot b + f \cdot d$		0.064***	0.178***	0.177***	0.178***	–0.019	–0.020	–0.005
		(0.025)	(0.037)	(0.037)	(0.038)	(0.029)	(0.030)	(0.028)
Geographic controls			✓	✓	✓	✓	✓	✓
Religious controls				✓	✓		✓	✓
Facilities controls					✓			✓
District fixed effects		✓	✓	✓	✓	✓	✓	✓
Observations	43,394	43,394	43,394	43,394	43,394	21,502	20,565	19,206

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if Golkar was the most voted party in the village in the 1999 parliamentary election and zero otherwise. Geographic controls include population density, a quartic in the logarithm of the village population, a quartic in the percentage of households whose main occupation is in agriculture, share of agricultural land in the village, distance to the subdistrict office, distance to the district capital, and indicators for urban and high altitude. Religious controls include the number of mosques, prayer houses, churches, and Buddhist temples per 1,000 people. Facilities controls include the number of hospitals, maternity hospitals, polyclinics, *puskesmas* (primary care centers), kindergartens, primary schools, high schools, and TVs per 1,000 people. Estimation of "*Kelurahan* indicator" (=OLS) is based on linear least squares regression, with controls for either the variables listed in the table (columns 1–5) or the propensity-score strata, province fixed effects, and the full set of interactions between the strata and the fixed effects (columns 6–8). In the latter case, the variables listed in the table correspond to the propensity score specifications. Estimation of ATT and ATC is described in Section 2 (in particular, see Theorem 1). Cluster-robust standard errors (columns 1–5, OLS), bootstrap standard errors (columns 6–8, OLS), and Huber–White standard errors (ATT, ATC, and ATE) are in parentheses. Huber–White standard errors ignore that the propensity score is estimated.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

and 0.788. Because—in this empirical context—we should arguably intend to estimate the average treatment effect, I also report a "properly reweighted" weighted average of $\hat{\tau}_{ATT}$ and $\hat{\tau}_{ATC}$, *i.e.* an estimate of the average effect of appointed officials (ATE). Since the weights underlying linear regression are poorly chosen, we can expect large differences between these estimates and the OLS estimates, and this is indeed the case. The results of

Table 6: Matching Estimates of the Effects of Local Officials on Electoral Results

	Linear probability model					Probit model		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ATT	—	0.007	0.070***	0.087***	0.069***	0.028*	0.030*	0.031*
		(0.008)	(0.025)	(0.025)	(0.026)	(0.016)	(0.016)	(0.016)
ATE	—	0.003	–0.019	–0.037	–0.003	–0.005	–0.007	–0.001
		(0.010)	(0.057)	(0.056)	(0.065)	(0.030)	(0.031)	(0.031)
Geographic controls			✓	✓	✓	✓	✓	✓
Religious controls				✓	✓		✓	✓
Facilities controls					✓			✓
District fixed effects		✓	✓	✓	✓	✓	✓	✓
Observations	43,394	43,394	43,394	43,394	43,394	21,502	20,565	19,206

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if Golkar was the most voted party in the village in the 1999 parliamentary election and zero otherwise. Geographic controls include population density, a quartic in the logarithm of the village population, a quartic in the percentage of households whose main occupation is in agriculture, share of agricultural land in the village, distance to the subdistrict office, distance to the district capital, and indicators for urban and high altitude. Religious controls include the number of mosques, prayer houses, churches, and Buddhist temples per 1,000 people. Facilities controls include the number of hospitals, maternity hospitals, polyclinics, *puskesmas* (primary care centers), kindergartens, primary schools, high schools, and TVs per 1,000 people. Estimation is based on nearest-neighbor matching on the estimated propensity score (with a single match). For columns 6–8, exact matching on province fixed effects is also required. The propensity score is estimated using a linear probability model (columns 1–5) or an algorithm based on a probit model (columns 6–8). A description of this algorithm is given in Martinez-Bravo (2014). Abadie–Imbens standard errors are in parentheses. These standard errors ignore that the propensity score is estimated.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

the decompositions in Table 5 are generally quite surprising, and they differ enormously between the linear probability model and the propensity score model. In the case of the linear probability model, all of the implied estimates of the average treatment effect are positive and statistically significant. The estimates from columns 3–5 are also very similar in magnitude, and they suggest that—on average—appointed officials increased the probability of Golkar victory by 18 percentage points. At the same time, however, the implied estimates of the average effect on *kelurahan* villages are close to zero and usually insignificant. When we turn to the results from the propensity score model, this pattern is reversed. The average effect of appointed officials on *kelurahan* villages seems to be relatively small in absolute value, but positive and significant; the average treatment effect is indistinguishable from zero.

Which of these patterns is believable? Is the average effect of appointed officials positive, but the average effect on *kelurahan* villages close to zero? Or, maybe the appointed officials increased the probability of Golkar victory only in the "treated" villages? Again, we might try to reconcile these conflicting findings using an alternative estimation

method. Therefore, Table 6 reports nearest-neighbor matching estimates of the average effect of appointed officials and of the average effect of appointed officials on *kelurahan* villages. The propensity score is estimated either using a linear probability model (as, implicitly, in Table 5) or using a specific algorithm based on a probit model (as, explicitly, in Martinez-Bravo, 2014). In the latter case, I also impose a requirement of exact matching within each province. As evident in Table 6, the pattern of estimated effects now becomes more coherent. The average effect on the "treated" seems to be positive and statistically significant; if we ignore column 2, the estimated effects vary between 3 and 9 percentage points. However, when we turn to the average effect of appointed officials, it is clear that all of the estimates are insignificant and very close to zero. Perhaps the average difference in electoral results between similar *kelurahan* and *desa* villages is actually negligible, which casts some doubt on one of the main conclusions in Martinez-Bravo (2014)—that the behavior of appointed and elected officials is, on average, very different.[10]

5 Summary

In this paper I study the interpretation of the least squares estimand in the homogeneous linear model when treatment effects are in fact heterogeneous. This problem is highly relevant for empirical economists, because many influential papers rely on linear least squares regression to provide estimates of the effects of various treatments, while it is clear that treatment effect heterogeneity is empirically important. How should we interpret the estimates in these studies? I derive a new theoretical result which demonstrates that linear least squares regression is equivalent to a weighted average of two estimators, both of which are likely to perform poorly in finite samples, with weights which are also poorly chosen. In particular, the weight which is placed by linear regression on the aver-

[10] Another conclusion in Martinez-Bravo (2014) is that the effect of appointed officials should be stronger in districts, in which Golkar was expected to win by a large margin, because such expectations incentivize these officials to manifest their allegiance to the regime. Also, the effect should be reversed in districts, in which PDI-P was expected to win by a large margin. These conclusions are tested in Appendix C, where various models are estimated on subsamples of the original data—and these subsamples are defined on the basis of district-level electoral results (PDI-P won large, PDI-P just won, Golkar just won, Golkar won large). Table C11 and Table C12 replicate the estimates from Martinez-Bravo (2014). Table C13 and Table C14 apply the main theoretical result of this paper to these estimates, and decompose all the coefficients from Table C11 and Table C12 into two components (ATT and ATC). Many of the results change. Table C15 and Table C16 present nearest-neighbor matching estimates of the average effect of appointed officials and of the average effect of appointed officials on *kelurahan* villages, separately for all of the subsamples. If we prefer the probit-based estimates of the propensity score and exact matching within each province, then this conclusion in Martinez-Bravo (2014) is correct for the effect of local officials on Golkar victory—this effect is positive only for districts, in which Golkar won by a large margin, and this includes the average treatment effect. However, when we turn to the effect on PDI-P victory, neither of the estimated effects is significantly different from zero—and they are usually very small.

age 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 average treatment effect on the treated. I also illustrate the importance of this result with two Monte Carlo studies, as well as with a replication of two prominent applied papers: Berger *et al.* (2013) on the effects of CIA interventions on international trade; and Martinez-Bravo (2014) on the effects of appointed officials on electoral outcomes. In both cases some important conclusions are not robust to allowing for heterogeneity in effects.

There are several lessons to be learned from this paper. First, empirical economists often believe that linear least squares regression provides a good approximation to the average treatment effect. Some authors only give their attention to issues of heterogeneity if this is motivated by a theoretical model or previous literature. However, linear regression might provide biased estimates of each of the relevant parameters of interest whenever heterogeneity is empirically important. Often, of course, this bias might be small, but this should never be taken for granted.

Second, it is useful to test for treatment effect heterogeneity. The main result of this paper (Theorem 1) provides a directly applicable decomposition for every least squares estimate, which can now be represented as a weighted average of two particular estimates: of the average treatment effect on the treated and of the average treatment effect on the controls. This decomposition can be applied as an easy-to-use informal test for treatment effect heterogeneity. However, more sophisticated procedures have also been developed, and can be used (see, *e.g.*, Crump, Hotz, Imbens and Mitnik, 2008).

Finally, it is essential to always define the parameter of interest. Many empirical papers lack a clear statement about the actual goal of the researcher—whether they are interested in the average effect, the average effect on some clearly defined population, or some other parameter. The linear regression estimand is never *the* most interesting parameter per se, and it might not correspond to any of the relevant parameters, as this paper also clarifies. Defining the parameter is important, because it enables the researcher to provide an interpretation of their result, and it also guarantees comparability between estimation methods. In some cases a precise definition of the parameter of interest might even allow the researcher to continue using linear least squares regression: as this paper clarifies, if nearly nobody gets treatment and we are interested in the effect on the treated, then we can maintain that we are approximately correct.

A Proofs

Proof of Theorem 1. First, consider $\mathrm{L}(y \mid 1, d, X) = \alpha + \tau d + X\beta$ (Equation 4). By the Frisch–Waugh theorem (Frisch and Waugh, 1933), $\tau = \tau_a$, where τ_a is defined by

$$\mathrm{L}[y \mid 1, d, p(X)] = \alpha_a + \tau_a d + \gamma_a \cdot p(X). \tag{12}$$

Second, notice that (12) is a linear projection of y on two variables: one binary and one continuous. We can therefore use the following result from Elder *et al.* (2010):

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

$$\begin{aligned} \tau_e &= \frac{\rho \cdot \mathrm{V}(x \mid d = 1)}{\rho \cdot \mathrm{V}(x \mid d = 1) + (1 - \rho) \cdot \mathrm{V}(x \mid d = 0)} \cdot w_1 \\ &+ \frac{(1 - \rho) \cdot \mathrm{V}(x \mid d = 0)}{\rho \cdot \mathrm{V}(x \mid d = 1) + (1 - \rho) \cdot \mathrm{V}(x \mid d = 0)} \cdot w_0, \end{aligned}$$

where

$$w_1 = \frac{\mathrm{Cov}(d, y)}{\mathrm{V}(d)} - \frac{\mathrm{Cov}(d, x)}{\mathrm{V}(d)} \cdot \frac{\mathrm{Cov}(x, y \mid d = 1)}{\mathrm{V}(x \mid d = 1)}$$

and

$$w_0 = \frac{\mathrm{Cov}(d, y)}{\mathrm{V}(d)} - \frac{\mathrm{Cov}(d, x)}{\mathrm{V}(d)} \cdot \frac{\mathrm{Cov}(x, y \mid d = 0)}{\mathrm{V}(x \mid d = 0)}.$$

Combining the two pieces gives

$$\begin{aligned} \tau &= \frac{\rho \cdot \mathrm{V}[p(X) \mid d = 1]}{\rho \cdot \mathrm{V}[p(X) \mid d = 1] + (1 - \rho) \cdot \mathrm{V}[p(X) \mid d = 0]} \cdot w_1^* \\ &+ \frac{(1 - \rho) \cdot \mathrm{V}[p(X) \mid d = 0]}{\rho \cdot \mathrm{V}[p(X) \mid d = 1] + (1 - \rho) \cdot \mathrm{V}[p(X) \mid d = 0]} \cdot w_0^*, \end{aligned} \tag{13}$$

where

$$w_1^* = \frac{\mathrm{Cov}(d, y)}{\mathrm{V}(d)} - \frac{\mathrm{Cov}[d, p(X)]}{\mathrm{V}(d)} \cdot \frac{\mathrm{Cov}[p(X), y \mid d = 1]}{\mathrm{V}[p(X) \mid d = 1]} \tag{14}$$

and

$$w_0^* = \frac{\mathrm{Cov}\,(d,y)}{\mathrm{V}\,(d)} - \frac{\mathrm{Cov}\,[d,p\,(X)]}{\mathrm{V}\,(d)} \cdot \frac{\mathrm{Cov}\,[p\,(X),y \mid d=0]}{\mathrm{V}\,[p\,(X) \mid d=0]}. \tag{15}$$

Third, notice that $w_1^* = \tau_{APE|d=0}$ and $w_0^* = \tau_{APE|d=1}$, as defined in (10). Indeed,

$$\frac{\mathrm{Cov}\,(d,y)}{\mathrm{V}\,(d)} = \mathrm{E}\,(y \mid d=1) - \mathrm{E}\,(y \mid d=0) \tag{16}$$

and also

$$\frac{\mathrm{Cov}\,[d,p\,(X)]}{\mathrm{V}\,(d)} = \mathrm{E}\,[p\,(X) \mid d=1] - \mathrm{E}\,[p\,(X) \mid d=0]\,. \tag{17}$$

Moreover, for $j = 0, 1$,

$$\frac{\mathrm{Cov}\,[p\,(X),y \mid d=j]}{\mathrm{V}\,[p\,(X) \mid d=j]} = \gamma_j \tag{18}$$

where γ_1 and γ_0 are defined in (7) and (8), respectively. Because

$$\begin{aligned} \mathrm{E}\,(y \mid d=1) - \mathrm{E}\,(y \mid d=0) &= (\mathrm{E}\,[p\,(X) \mid d=1] - \mathrm{E}\,[p\,(X) \mid d=0]) \cdot \gamma_1 \\ &+ (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot \mathrm{E}\,[p\,(X) \mid d=0] \end{aligned} \tag{19}$$

and also

$$\begin{aligned} \mathrm{E}\,(y \mid d=1) - \mathrm{E}\,(y \mid d=0) &= (\mathrm{E}\,[p\,(X) \mid d=1] - \mathrm{E}\,[p\,(X) \mid d=0]) \cdot \gamma_0 \\ &+ (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot \mathrm{E}\,[p\,(X) \mid d=1]\,, \end{aligned} \tag{20}$$

where again α_1 and α_0 are defined in (7) and (8), we get the result that $w_1^* = \tau_{APE|d=0}$ and $w_0^* = \tau_{APE|d=1}$. Interestingly, Equations 19 and 20 are special cases of the Oaxaca–Blinder decomposition (Blinder, 1973; Oaxaca, 1973; Fortin, Lemieux and Firpo, 2011).

Combining the three pieces gives

$$\begin{aligned} \tau &= \frac{\rho \cdot \mathrm{V}\,[p\,(X) \mid d=1]}{\rho \cdot \mathrm{V}\,[p\,(X) \mid d=1] + (1-\rho) \cdot \mathrm{V}\,[p\,(X) \mid d=0]} \cdot \tau_{APE|d=0} \\ &+ \frac{(1-\rho) \cdot \mathrm{V}\,[p\,(X) \mid d=0]}{\rho \cdot \mathrm{V}\,[p\,(X) \mid d=1] + (1-\rho) \cdot \mathrm{V}\,[p\,(X) \mid d=0]} \cdot \tau_{APE|d=1}, \end{aligned} \tag{21}$$

which completes the proof. □

B Further Monte Carlo Results

Figure B3: Linear Regression and LPM-Based Estimates of Average Treatment Effects

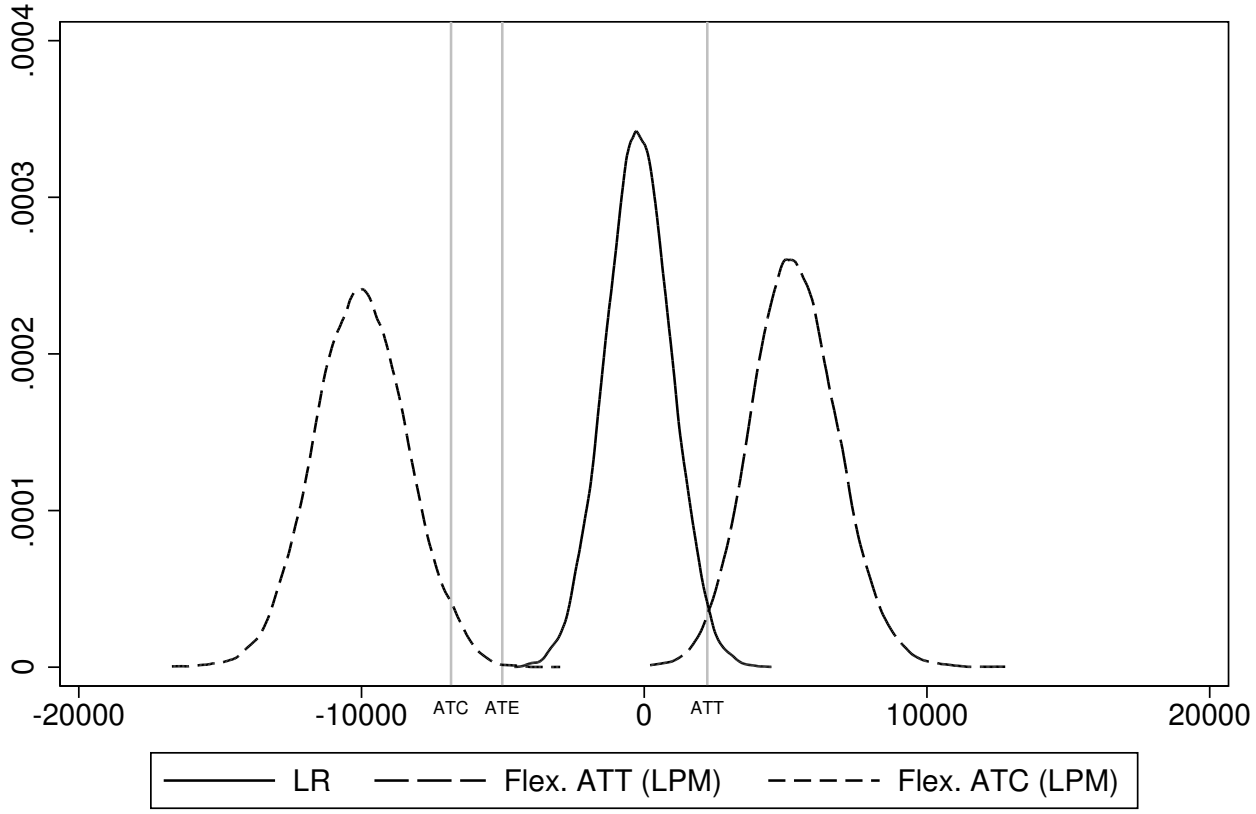


Table B7: Simulation Results of the First MC Study

Method	Parameter	Mean bias	Median bias	RMSE	MAE	SD
LR	ATE	4812	4810	4952	4810	1170
LR	ATT	–2439	–2441	2705	2441	1170
LR	ATC	6625	6623	6727	6623	1170
Flex. ATE	ATE	40	22	2601	1726	2601
Flex. ATT	ATT	–125	–133	1368	922	1362
Flex. ATC	ATC	76	42	3188	2114	3187
Flex. ATT (LPM)	ATT	3093	3047	3454	3047	1537
Flex. ATC (LPM)	ATC	–3193	–3201	3590	3201	1642

Notes: "Method" refers to the estimation method. "Parameter" refers to the parameter of interest, against which biases are calculated. "LR" denotes linear least squares regression. "Flex. ATE", "Flex. ATT", and "Flex. ATC" denote various versions of the "flexible OLS" estimator. "Flex. ATT (LPM)" and "Flex. ATC (LPM)" denote various versions of the LPM-based "flexible OLS" estimator. See the text for details.

Table B8: Implicit Weights on $\hat{\tau}_{ATT}$ and $\hat{\tau}_{ATC}$ in the First MC Study

	Mean	SD	Minimum	Maximum
$\mathrm{P}(d=1)$	0.200	0.014	0.146	0.260
$\mathrm{P}(d=0)$	0.800	0.014	0.740	0.854
w_{ATT}	0.640	0.038	0.457	0.767
w_{ATC}	0.360	0.038	0.233	0.543

Notes: $\mathrm{P}(d=1)$ denotes the proportion of treated units. $\mathrm{P}(d=0)$ denotes the proportion of control units. The implicit weights on $\hat{\tau}_{ATT}$ and $\hat{\tau}_{ATC}$ are denoted by w_{ATT} and w_{ATC}, respectively.

Table B9: Simulation Results of the Second MC Study

$\mathrm{P}(d=1)$	Mean bias	Median bias	RMSE	MAE	SD
			Panel A: ATT		
0.200	–2304	–2273	2398	2273	662
0.221	–2485	–2457	2577	2457	685
0.247	–2638	–2622	2735	2622	722
0.281	–2881	–2849	2985	2849	781
0.324	–3100	–3030	3225	3030	888
0.384	–3399	–3318	3535	3318	971
0.471	–3761	–3647	3926	3647	1128
0.609	–4380	–4195	4607	4195	1428
0.862	–5962	–5647	6411	5647	2357
			Panel B: ATC		
0.200	6759	6791	6791	6791	662
0.221	6579	6606	6614	6606	685
0.247	6426	6442	6466	6442	722
0.281	6182	6214	6232	6214	781
0.324	5963	6034	6029	6034	888
0.384	5665	5745	5747	5745	971
0.471	5303	5416	5421	5416	1128
0.609	4683	4869	4896	4869	1428
0.862	3101	3416	3895	3528	2357

Notes: $\mathrm{P}(d=1)$ denotes the proportion of treated units. Simulation results are reported for linear least squares regression. Biases are calculated against either the average treatment effect on the treated (Panel A) or the average treatment effect on the controls (Panel B).

Table B10: Implicit Weights on $\hat{\tau}_{ATT}$ and $\hat{\tau}_{ATC}$ in the Second MC Study

	w_{ATT}				w_{ATC}			
P $(d = 1)$	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
0.200	0.638	0.032	0.525	0.722	0.362	0.032	0.278	0.475
0.221	0.618	0.035	0.485	0.726	0.382	0.035	0.274	0.515
0.247	0.598	0.036	0.456	0.705	0.402	0.036	0.295	0.544
0.281	0.575	0.039	0.434	0.689	0.425	0.039	0.311	0.566
0.324	0.551	0.041	0.376	0.674	0.449	0.041	0.326	0.624
0.384	0.523	0.042	0.364	0.636	0.477	0.042	0.364	0.636
0.471	0.494	0.045	0.351	0.647	0.506	0.045	0.353	0.649
0.609	0.459	0.048	0.321	0.619	0.541	0.048	0.381	0.679
0.862	0.368	0.072	0.128	0.593	0.632	0.072	0.407	0.872

Notes: P $(d = 1)$ denotes the proportion of treated units. The implicit weights on $\hat{\tau}_{ATT}$ and $\hat{\tau}_{ATC}$ are denoted by w_{ATT} and w_{ATC}, respectively.

C Further Results on the Effects of Local Officials

Table C11: A Replication of Martinez-Bravo (2014)—The Effects on Golkar Victory

	Whole sample	PDI-P won large 1999	PDI-P just won 1999	Golkar just won 1999	Golkar won large 1999	Neither won
			Linear probability model			
Kelurahan indicator	0.055***	0.002	0.076**	0.128***	0.044**	0.068*
	(0.012)	(0.016)	(0.029)	(0.037)	(0.018)	(0.038)
Observations	43,394	15,430	9,114	5,946	7,378	5,526
			Propensity score model			
Kelurahan indicator	0.033***	0.001	0.034***	0.136***	0.047***	0.028
	(0.009)	(0.006)	(0.010)	(0.050)	(0.016)	(0.025)
Observations	19,206	7,814	4,303	1,822	3,378	1,889

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if Golkar was the most voted party in the village in the 1999 parliamentary election and zero otherwise. All regressions include geographic controls, religious controls, facilities controls, and district fixed effects. Estimation is based on linear least squares regression, with controls for either the variables listed in the table ("Linear probability model") or the propensity-score strata, province fixed effects, and the full set of interactions between the strata and the fixed effects ("Propensity score model"). In the latter case, the variables listed in the table correspond to the propensity score specifications. Cluster-robust standard errors ("Linear probability model") and bootstrap standard errors ("Propensity score model") are in parentheses.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table C12: A Replication of Martinez-Bravo (2014)—The Effects on PDI-P Victory

	Whole sample	PDI-P won large 1999	PDI-P just won 1999	Golkar just won 1999	Golkar won large 1999	Neither won
			Linear probability model			
Kelurahan indicator	–0.021	0.037*	–0.037	–0.087*	–0.024	–0.004
	(0.014)	(0.021)	(0.045)	(0.043)	(0.015)	(0.045)
Observations	43,394	15,430	9,114	5,946	7,378	5,526
			Propensity score model			
Kelurahan indicator	–0.003	0.033***	–0.008	–0.099***	–0.021*	–0.023
	(0.010)	(0.009)	(0.039)	(0.036)	(0.011)	(0.045)
Observations	19,206	7,814	4,303	1,822	3,378	1,889

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if PDI-P was the most voted party in the village in the 1999 parliamentary election and zero otherwise. All regressions include geographic controls, religious controls, facilities controls, and district fixed effects. Estimation is based on linear least squares regression, with controls for either the variables listed in the table ("Linear probability model") or the propensity-score strata, province fixed effects, and the full set of interactions between the strata and the fixed effects ("Propensity score model"). In the latter case, the variables listed in the table correspond to the propensity score specifications. Cluster-robust standard errors ("Linear probability model") and bootstrap standard errors ("Propensity score model") are in parentheses.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table C13: Martinez-Bravo (2014) and Treatment Effect Heterogeneity—The Effects on Golkar Victory

	Whole sample	PDI-P won large 1999	PDI-P just won 1999	Golkar just won 1999	Golkar won large 1999	Neither won
			Linear probability model			
Kelurahan indicator	0.055***	0.002	0.076**	0.128***	0.044**	0.068*
	(0.012)	(0.016)	(0.029)	(0.037)	(0.018)	(0.038)
Decomposition (Theorem 1)						
a. ATT	–0.009	0.016	0.107***	0.094**	0.012	0.093*
	(0.028)	(0.032)	(0.039)	(0.047)	(0.027)	(0.051)
b. w_{ATT}	0.679	0.614	0.715	0.586	0.603	0.771
c. ATC	0.192***	–0.022	–0.001	0.175***	0.093***	–0.017
	(0.042)	(0.017)	(0.065)	(0.063)	(0.025)	(0.054)
d. w_{ATC}	0.321	0.386	0.285	0.414	0.397	0.229
OLS $= a \cdot b + c \cdot d$	0.055***	0.002	0.076**	0.128***	0.044**	0.068*
	(0.012)	(0.016)	(0.029)	(0.037)	(0.018)	(0.038)
e. P $(d = 1)$	0.070	0.070	0.060	0.060	0.110	0.045
f. P $(d = 0)$	0.930	0.930	0.940	0.940	0.890	0.955
ATE $= e \cdot b + f \cdot d$	0.178***	–0.019	0.006	0.170***	0.084***	–0.012
	(0.038)	(0.016)	(0.061)	(0.060)	(0.022)	(0.052)
Observations	43,394	15,430	9,114	5,946	7,378	5,526
			Propensity score model			
Kelurahan indicator	0.033***	0.001	0.034***	0.136***	0.047***	0.028
	(0.009)	(0.006)	(0.010)	(0.050)	(0.016)	(0.025)
Decomposition (Theorem 1)						
a. ATT	0.045***	0.004	0.064*	0.100	0.048**	0.034
	(0.016)	(0.011)	(0.034)	(0.061)	(0.022)	(0.028)
b. w_{ATT}	0.779	0.852	0.782	0.705	0.660	0.879
c. ATC	–0.011	–0.011	–0.073	0.224***	0.045*	–0.019
	(0.032)	(0.019)	(0.063)	(0.071)	(0.023)	(0.032)
d. w_{ATC}	0.221	0.148	0.218	0.295	0.340	0.121
OLS $= a \cdot b + c \cdot d$	0.033***	0.001	0.034***	0.136***	0.047***	0.028
	(0.009)	(0.006)	(0.010)	(0.050)	(0.016)	(0.025)
e. P $(d = 1)$	0.116	0.099	0.104	0.110	0.181	0.100
f. P $(d = 0)$	0.884	0.901	0.896	0.890	0.819	0.900
ATE $= e \cdot b + f \cdot d$	–0.005	–0.009	–0.059	0.210***	0.046**	–0.013
	(0.028)	(0.017)	(0.058)	(0.067)	(0.021)	(0.029)
Observations	19,206	7,814	4,303	1,822	3,378	1,889

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if Golkar was the most voted party in the village in the 1999 parliamentary election and zero otherwise. All regressions and propensity score specifications include geographic controls, religious controls, facilities controls, and district fixed effects. Estimation of "*Kelurahan* indicator" (=OLS) is based on linear least squares regression, with controls for either the variables listed in the table ("Linear probability model") or the propensity-score strata, province fixed effects, and the full set of interactions between the strata and the fixed effects ("Propensity score model"). In the latter case, the variables listed in the table correspond to the propensity score specifications. Estimation of ATT and ATC is described in Section 2 (in particular, see Theorem 1). Cluster-robust standard errors ("Linear probability model", OLS), bootstrap standard errors ("Propensity score model", OLS), and Huber–White standard errors (ATT, ATC, and ATE) are in parentheses. Huber–White standard errors ignore that the propensity score is estimated.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table C14: Martinez-Bravo (2014) and Treatment Effect Heterogeneity—The Effects on PDI-P Victory

	Whole sample	PDI-P won large 1999	PDI-P just won 1999	Golkar just won 1999	Golkar won large 1999	Neither won
			Linear probability model			
Kelurahan indicator	–0.021	0.037*	–0.037	–0.087*	–0.024	–0.004
	(0.014)	(0.021)	(0.045)	(0.043)	(0.015)	(0.045)
Decomposition (Theorem 1)						
a. ATT	0.012	0.006	–0.069	–0.110	0.011	–0.024
	(0.025)	(0.040)	(0.059)	(0.074)	(0.024)	(0.051)
b. w_{ATT}	0.679	0.614	0.715	0.586	0.603	0.771
c. ATC	–0.091**	0.086***	0.043	–0.055	–0.078***	0.065
	(0.041)	(0.025)	(0.062)	(0.047)	(0.021)	(0.057)
d. w_{ATC}	0.321	0.386	0.285	0.414	0.397	0.229
OLS $= a \cdot b + c \cdot d$	–0.021	0.037*	–0.037	–0.087*	–0.024	–0.004
	(0.014)	(0.021)	(0.045)	(0.043)	(0.015)	(0.045)
e. P $(d = 1)$	0.070	0.070	0.060	0.060	0.110	0.045
f. P $(d = 0)$	0.930	0.930	0.940	0.940	0.890	0.955
ATE $= e \cdot b + f \cdot d$	–0.084**	0.080***	0.036	–0.058	–0.068***	0.061
	(0.037)	(0.023)	(0.059)	(0.046)	(0.019)	(0.056)
Observations	43,394	15,430	9,114	5,946	7,378	5,526
			Propensity score model			
Kelurahan indicator	–0.003	0.033***	–0.008	–0.099***	–0.021*	–0.023
	(0.010)	(0.009)	(0.039)	(0.036)	(0.011)	(0.045)
Decomposition (Theorem 1)						
a. ATT	–0.025	0.029	–0.030	–0.098	–0.021	–0.030
	(0.020)	(0.021)	(0.047)	(0.061)	(0.016)	(0.055)
b. w_{ATT}	0.779	0.852	0.782	0.705	0.660	0.879
c. ATC	0.073**	0.054*	0.070	–0.102	–0.020	0.032
	(0.032)	(0.031)	(0.066)	(0.076)	(0.021)	(0.064)
d. w_{ATC}	0.221	0.148	0.218	0.295	0.340	0.121
OLS $= a \cdot b + c \cdot d$	–0.003	0.033***	–0.008	–0.099***	–0.021*	–0.023
	(0.010)	(0.009)	(0.039)	(0.036)	(0.011)	(0.045)
e. P $(d = 1)$	0.116	0.099	0.104	0.110	0.181	0.100
f. P $(d = 0)$	0.884	0.901	0.896	0.890	0.819	0.900
ATE $= e \cdot b + f \cdot d$	0.062**	0.052*	0.059	–0.102	–0.020	0.026
	(0.029)	(0.028)	(0.062)	(0.073)	(0.019)	(0.060)
Observations	19,206	7,814	4,303	1,822	3,378	1,889

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if PDI-P was the most voted party in the village in the 1999 parliamentary election and zero otherwise. All regressions and propensity score specifications include geographic controls, religious controls, facilities controls, and district fixed effects. Estimation of "*Kelurahan* indicator" (=OLS) is based on linear least squares regression, with controls for either the variables listed in the table ("Linear probability model") or the propensity-score strata, province fixed effects, and the full set of interactions between the strata and the fixed effects ("Propensity score model"). In the latter case, the variables listed in the table correspond to the propensity score specifications. Estimation of ATT and ATC is described in Section 2 (in particular, see Theorem 1). Cluster-robust standard errors ("Linear probability model", OLS), bootstrap standard errors ("Propensity score model", OLS), and Huber–White standard errors (ATT, ATC, and ATE) are in parentheses. Huber–White standard errors ignore that the propensity score is estimated.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table C15: Matching Estimates of the Effects of Local Officials on Golkar Victory

	Whole sample	PDI-P won large 1999	PDI-P just won 1999	Golkar just won 1999	Golkar won large 1999	Neither won
ATT-LPM	0.069***	–0.021	0.087**	–0.031	0.115***	0.032
	(0.026)	(0.041)	(0.039)	(0.129)	(0.036)	(0.041)
ATE-LPM	–0.003	–0.062	0.307*	–0.151	0.116	0.570**
	(0.065)	(0.049)	(0.163)	(0.207)	(0.081)	(0.266)
Observations	43,394	15,430	9,114	5,946	7,378	5,526
ATT-probit	0.031*	–0.006	0.000	0.104	0.044*	0.016
	(0.016)	(0.022)	(0.042)	(0.081)	(0.026)	(0.049)
ATE-probit	–0.001	–0.008	0.037	0.131	0.070**	–0.037
	(0.031)	(0.044)	(0.076)	(0.142)	(0.032)	(0.069)
Observations	19,206	7,814	4,303	1,822	3,378	1,889

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if Golkar was the most voted party in the village in the 1999 parliamentary election and zero otherwise. All propensity score specifications include geographic controls, religious controls, facilities controls, and district fixed effects. Estimation is based on nearest-neighbor matching on the estimated propensity score (with a single match). For "ATT-probit" and "ATE-probit", exact matching on province fixed effects is also required. The propensity score is estimated using a linear probability model ("ATT-LPM" and "ATE-LPM") or an algorithm based on a probit model ("ATT-probit" and "ATE-probit"). A description of this algorithm is given in Martinez-Bravo (2014). Abadie–Imbens standard errors are in parentheses. These standard errors ignore that the propensity score is estimated.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table C16: Matching Estimates of the Effects of Local Officials on PDI-P Victory

	Whole sample	PDI-P won large 1999	PDI-P just won 1999	Golkar just won 1999	Golkar won large 1999	Neither won
ATT-LPM	–0.015	0.047	–0.060	0.059	–0.075**	–0.004
	(0.027)	(0.051)	(0.049)	(0.119)	(0.030)	(0.058)
ATE-LPM	0.120*	0.118*	–0.142	–0.145	–0.074	–0.117
	(0.070)	(0.063)	(0.176)	(0.156)	(0.070)	(0.217)
Observations	43,394	15,430	9,114	5,946	7,378	5,526
ATT-probit	–0.008	0.043	–0.034	–0.055	–0.015	–0.037
	(0.020)	(0.030)	(0.053)	(0.078)	(0.020)	(0.092)
ATE-probit	0.014	0.037	0.122	–0.087	–0.033	–0.107
	(0.041)	(0.058)	(0.099)	(0.123)	(0.026)	(0.122)
Observations	19,206	7,814	4,303	1,822	3,378	1,889

Notes: See also Martinez-Bravo (2014) for more details on these data. The unit of observation is a village. The dependent variable equals one if PDI-P was the most voted party in the village in the 1999 parliamentary election and zero otherwise. All propensity score specifications include geographic controls, religious controls, facilities controls, and district fixed effects. Estimation is based on nearest-neighbor matching on the estimated propensity score (with a single match). For "ATT-probit" and "ATE-probit", exact matching on province fixed effects is also required. The propensity score is estimated using a linear probability model ("ATT-LPM" and "ATE-LPM") or an algorithm based on a probit model ("ATT-probit" and "ATE-probit"). A description of this algorithm is given in Martinez-Bravo (2014). Abadie–Imbens standard errors are in parentheses. These standard errors ignore that the propensity score is estimated.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

References

ALESINA, A., GIULIANO, P. & NUNN, N. (2013). On the origins of gender roles: Women and the plough. *Quarterly Journal of Economics* 128, 469–530.

ALMOND, D., CHAY, K. Y. & LEE, D. S. (2005). The costs of low birth weight. *Quarterly Journal of Economics* 120, 1031–1083.

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., GRADDY, K. & IMBENS, G. W. (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. & KRUEGER, A. B. (1999). Empirical strategies in labor economics. In: *Handbook of Labor Economics* (ASHENFELTER, O. & CARD, D., eds.), vol. 3A. North-Holland.

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

ARONOW, P. M. & SAMII, C. (2014). Does regression produce representative estimates of causal effects? Unpublished.

BERGER, D., EASTERLY, W., NUNN, N. & SATYANATH, S. (2013). Commercial imperialism? Political influence and trade during the Cold War. *American Economic Review* 103, 863–896.

BITLER, M. P., GELBACH, J. B. & HOYNES, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review* 96, 988–1012.

BITLER, M. P., GELBACH, J. B. & HOYNES, H. W. (2008). Distributional impacts of the Self-Sufficiency Project. *Journal of Public Economics* 92, 748–765.

BLACK, D. A., SMITH, J. A., BERGER, M. C. & NOEL, B. J. (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.

BOND, T. N. & LANG, K. (2013). The evolution of the black-white test score gap in grades K–3: The fragility of results. *Review of Economics and Statistics* 95, 1468–1479.

BOUSTAN, L. P. & COLLINS, W. J. (2013). The origins and persistence of black-white differences in women's labor force participation. NBER Working Paper no. 19040.

BUSSO, M., DINARDO, J. & MCCRARY, J. (2013). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* (forthcoming).

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I., HAHN, J. & NEWEY, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica* 81, 535–580.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. & MITNIK, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics* 90, 389–405.

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

DEATON, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature* 48, 424–455.

DIETERLE, S. & SNELL, A. (2014). Exploiting nonlinearities in the first stage regressions of IV procedures. Unpublished.

ELDER, T. E., GODDEERIS, J. H. & HAIDER, S. J. (2010). Unexplained gaps and Oaxaca–Blinder decompositions. *Labour Economics* 17, 284–290.

FORTIN, N., LEMIEUX, T. & FIRPO, S. (2011). Decomposition methods in economics. In: *Handbook of Labor Economics* (ASHENFELTER, O. & CARD, D., eds.), vol. 4A. North-Holland.

FREEDMAN, D. A. (2008a). On regression adjustments in experiments with several treatments. *Annals of Applied Statistics* 2, 176–196.

FREEDMAN, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics* 40, 180–193.

FRISCH, R. & WAUGH, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica* 1, 387–401.

FRYER, R. G. & GREENSTONE, M. (2010). The changing consequences of attending historically black colleges and universities. *American Economic Journal: Applied Economics* 2, 116–148.

FRYER, R. G. & LEVITT, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics* 86, 447–464.

FRYER, R. G. & LEVITT, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics* 2, 210–240.

GIBBONS, C. E., SUÁREZ SERRATO, J. C. & URBANCIC, M. B. (2014). Broken or fixed effects? NBER Working Paper no. 20342.

GITTLEMAN, M. & WOLFF, E. N. (2004). Racial differences in patterns of wealth accumulation. *Journal of Human Resources* 39, 193–227.

HECKMAN, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture. *Journal of Political Economy* 109, 673–748.

HECKMAN, J. J., URZUA, S. & VYTLACIL, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 88, 389–432.

HECKMAN, J. J. & VYTLACIL, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73, 669–738.

HECKMAN, J. J. & VYTLACIL, E. J. (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* (HECKMAN, J. J. & LEAMER, E. E., eds.), vol. 6B. North-Holland.

HOLLAND, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–960.

HUMPHREYS, M. (2009). Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities. Unpublished.

IMAI, K. & KIM, I. S. (2013). On the use of linear fixed effects regression estimators for causal inference. Unpublished.

IMBENS, G. W. (2014). Matching methods in practice: Three examples. NBER Working Paper no. 19959.

IMBENS, G. W. & ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* 62, 467–475.

IMBENS, G. W. & WOOLDRIDGE, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47, 5–86.

KHWAJA, A., PICONE, G., SALM, M. & TROGDON, J. G. (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.

KLINE, P. (2014). A note on variance estimation for the Oaxaca estimator of average treatment effects. *Economics Letters* 122, 428–431.

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.

LANG, K. & MANOVE, M. (2011). Education and labor market discrimination. *American Economic Review* 101, 1467–1496.

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., MOGSTAD, M. & WISWALL, M. (2012). What linear estimators miss: The effects of family income on child outcomes. *American Economic Journal: Applied Economics* 4, 1–35.

MARTINEZ-BRAVO, M. (2014). The role of local officials in new democracies: Evidence from Indonesia. *American Economic Review* 104, 1244–1287.

OAXACA, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review* 14, 693–709.

RHODES, W. (2010). Heterogeneous treatment effects: What does a regression estimate? *Evaluation Review* 34, 334–361.

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

ROTHSTEIN, J. & WOZNY, N. (2013). Permanent income and the black-white test score gap. *Journal of Human Resources* 48, 509–544.

SCHOCHET, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *Journal of Statistical Planning and Inference* 140, 246–259.

SOLON, G., HAIDER, S. J. & WOOLDRIDGE, J. (2013). What are we weighting for? NBER Working Paper no. 18859.

VOGL, T. S. (2013). Marriage institutions and sibling competition: Evidence from South Asia. *Quarterly Journal of Economics* 128, 1017–1072.

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.

WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd ed.

YITZHAKI, S. (1996). On using linear regressions in welfare economics. *Journal of Business & Economic Statistics* 14, 478–486.