When Should We (Not) Interpret Linear IV Estimands as LATE?

Tymon Słoczyński

Abstract

In this paper I revisit the interpretation of the linear instrumental variables (IV) estimand as a weighted average of conditional local average treatment effects (LATEs). I focus on a practically relevant situation in which additional covariates are required for identification but the reduced-form and first-stage regressions are possibly misspecified as a result of neglected heterogeneity in the effects of the instrument. If we also allow for conditional monotonicity, *i.e.* the existence of compliers but no defiers at some covariate values and the existence of defiers but no compliers elsewhere, then the weights on some conditional LATEs are negative and the IV estimand is no longer interpretable as a causal effect. Even if monotonicity holds unconditionally, the IV estimand is not interpretable as the unconditional LATE parameter unless the groups that are encouraged and not encouraged to get treated are roughly equal sized.

Tymon Słoczyński is an Assistant Professor at the Department of Economics and International Business School, Brandeis University. E-mail: tslocz@brandeis.edu.

This version: November 12, 2020. Section 3.4 of this paper draws and expands some of the materials from Słoczyński (2018, Section 5). I am grateful to Pedro Sant'Anna for helpful comments. I also thank Anne Laski for excellent research assistance.

1 Introduction

Many instrumental variables are only valid after conditioning on additional covariates. Angrist and Imbens (1995) provide an influential interpretation of the two-stage least squares (2SLS) estimand in this context as a convex combination of conditional local average treatment effects (LATEs), *i.e.* average effects of treatment for individuals whose treatment status is affected by the instrument. However, Angrist and Imbens (1995) restrict their attention to saturated models with discrete covariates and implicitly require that the researcher estimates a separate first-stage coefficient on the instrument for every combination of covariate values. Such restrictions are largely absent from empirical work, which limits the applicability of this result to interpreting actual IV estimates.

Recent contributions to this line of research, notably those of Kolesár (2013) and Evdokimov and Kolesár (2019), relax many of the practical limitations of Angrist and Imbens (1995)'s result and support the view that linear IV and 2SLS estimands can generally be written as a convex combination of conditional LATEs. However, Evdokimov and Kolesár (2019) assume that the reduced-form and first-stage regressions are correctly specified, which is perhaps implausible in the usual application of instrumental variables that limits the effects of the instrument in these regressions to be homogeneous. The results in Kolesár (2013) are more general and allow for misspecification. Also, unlike in most previous studies, Kolesár (2013) allows for conditional monotonicity, *i.e.* the existence of compliers but no defiers at some covariate values and the existence of defiers but no compliers elsewhere.[1] Still, Kolesár (2013) concludes that the interpretation of linear IV and 2SLS estimands as a convex combination of conditional LATEs is generally correct "under some mild assumptions about the first stage."

In this paper I present a more pessimistic view of the causal interpretability of linear IV and 2SLS estimands. In particular, I make three main contributions to the literature on IV regression.

[1]Following Angrist, Imbens, and Rubin (1996), "compliers" are individuals who get treated when encouraged to do so but not otherwise, while "defiers" are those who do not get treated when encouraged to do so and get treated otherwise. Usually, the existence of defiers is ruled out for all covariate values (*e.g.*, Abadie, 2003; Frölich, 2007), and the instrument is assumed to influence treatment status in only one direction. In a recent paper, Semenova (2020) argues that unconditional monotonicity is often implausible and provides evidence against this assumption in a randomized evaluation of the Job Corps training program (Schochet, Burghardt, and McConnell, 2008; Lee, 2009).

First, I demonstrate that when IV is applied in the usual way, with homogeneous effects of the instrument in the reduced-form and first-stage regressions, the weights on some conditional LATEs are negative under Kolesár (2013)'s assumptions. It follows that the IV estimand may no longer be interpretable as a causal effect; this parameter may turn out to be negative (positive) even if treatment effects are positive (negative) for everyone in the population.

Second, unlike in previous contributions to this literature, I explicitly compare the weights in the usual application of IV and in the overidentified specification of Angrist and Imbens (1995) with the "desired" weights that recover the unconditional LATE parameter. The advantage of Angrist and Imbens (1995)'s specification is that it guarantees to produce a convex combination of conditional LATEs even under conditional monotonicity. However, if monotonicity holds unconditionally, the difference between the "desired" weights and Angrist and Imbens (1995)'s weights is greater than that between the "desired" weights and the weights in the standard (just identified) specification. It turns out that, under unconditional monotonicity, if we misspecify the structural model, it may be desirable to also misspecify the first stage.

Finally, I demonstrate that the weights in the standard specification are often problematic for interpretation even under unconditional monotonicity. I show that the IV estimand may be quantitatively and qualitatively different from the unconditional LATE parameter whenever the groups that are encouraged and not encouraged to get treated (*i.e.* with different values of the instrument) are not approximately equal sized. Put another way, I demonstrate that the IV estimand may be substantially different from the parameter of interest even if all weights are positive and integrate to one, unless the relevant population is balanced in a particular sense.

2 Framework

In this section I formally define the statistical objects of interest, *i.e.* the conditional and unconditional IV and 2SLS estimands. I also review identification in the LATE framework with covariates, as previously discussed by Abadie (2003) and Frölich (2007), among others. Unlike in most pre-

vious studies, I devote particular attention to the possibility that the monotonicity assumption is valid only conditional on covariates (see also Kolesár, 2013; Semenova, 2020). Throughout the paper I also assume that the appropriate moments exist whenever necessary.

2.1 Notation and Estimands

Suppose that we are interested in the causal effect of a binary treatment, D, on an outcome, Y. For every individual, we define two potential outcomes, $Y(1)$ and $Y(0)$, which correspond to the values of Y that this individual would attain if treated ($D = 1$) and if not treated ($D = 0$), respectively. Thus, $Y(1) - Y(0)$ is the treatment effect. The treatment D is allowed to be endogenous but a binary instrument, Z, is also available. Let $D(1)$ and $D(0)$ denote the potential treatment statuses that correspond to the treatment actually received by an individual when she is encouraged ($Z = 1$) and not encouraged ($Z = 0$) to get treated, respectively. Consequently, $Y = Y(D)$ and $D = D(Z)$. If the observed outcome were to depend directly on Z, we would write $Y = Y(Z, D)$. Finally, let $X = (1, X_1, \ldots, X_J)$ denote a row vector of covariates. In some cases I will allow for the possibility that additional instruments have been created by interacting Z with all elements of X; then, $Z_C = (Z, ZX_1, \ldots, ZX_J)$ will be used to denote the resulting row vector of instruments.

To provide motivation for what follows, let us consider the standard single-equation linear model for our outcome of interest:

$$Y = \beta D + X\gamma + \upsilon, \tag{1}$$

where X and the instrument(s) are assumed to be uncorrelated with υ. Also, β is the main coefficient of interest. Unlike in textbook treatments of this model but in line with the literature on local average treatment effects, I do not assume that equation (1), often referred to as the "structural model," is correctly specified; in particular, I allow the effect of D on Y to be heterogeneous and correlated with both observables and unobservables.

In practice, however, many researchers act as if this model is correctly specified and use linear IV or 2SLS for estimation. In what follows, I will focus on the interpretation of the probability

limits of the IV and 2SLS estimators of β when the structural model is possibly misspecified. With a single instrument, the probability limit of linear IV or, simply, the (linear) IV estimand is

$$\beta_{\text{IV}} = \left[\left(\mathrm{E}\left[Q'W \right] \right)^{-1} \mathrm{E}\left[Q'Y \right] \right]_1, \tag{2}$$

where $W = (D, X)$, $Q = (Z, X)$, and $[\cdot]_k$ denotes the kth element of the corresponding vector. It is useful to note that equation (2) characterizes the usual (just identified) application of instrumental variables when a single instrument is available. This specification also corresponds to reduced-form and first-stage regressions that are separable in X and Z, and limit the effects of Z on Y and D to be homogeneous.

If a vector of instruments, Z_{C}, has been created and 2SLS is used for estimation, the relevant probability limit or, simply, the 2SLS estimand is

$$\beta_{\text{2SLS}} = \left[\left(\mathrm{E}\left[W'Q_{\text{C}} \right] \left(\mathrm{E}\left[Q_{\text{C}}'Q_{\text{C}} \right] \right)^{-1} \mathrm{E}\left[Q_{\text{C}}'W \right] \right)^{-1} \mathrm{E}\left[W'Q_{\text{C}} \right] \left(\mathrm{E}\left[Q_{\text{C}}'Q_{\text{C}} \right] \right)^{-1} \mathrm{E}\left[Q_{\text{C}}'Y \right] \right]_1, \tag{3}$$

where $Q_{\text{C}} = (Z_{\text{C}}, X)$. In this specification, the corresponding reduced-form and first-stage regressions are no longer separable in X and Z, and hence we implicitly allow for heterogeneity in the effects of Z on Y and D.

Regardless of the implicit restrictions on the effects of the instrument, the true first stage can generally be written as

$$\mathrm{E}\left[D \mid X, Z \right] = \psi(X) + \omega(X) \cdot Z, \tag{4}$$

where

$$\omega(x) = \mathrm{E}\left[D \mid Z = 1, X = x \right] - \mathrm{E}\left[D \mid Z = 0, X = x \right] \tag{5}$$

is the conditional first-stage slope coefficient or, equivalently, the coefficient on Z in the regression of D on 1 and Z in the subpopulation with $X = x$. Similarly, the conditional IV (or Wald) estimand can be written as

$$\beta(x) = \frac{\mathrm{E}\left[Y \mid Z = 1, X = x \right] - \mathrm{E}\left[Y \mid Z = 0, X = x \right]}{\mathrm{E}\left[D \mid Z = 1, X = x \right] - \mathrm{E}\left[D \mid Z = 0, X = x \right]}. \tag{6}$$

This parameter is equivalent to the coefficient on D in the IV regression of Y on 1 and D in the subpopulation with $X = x$, with Z as the instrument for D.

2.2 Local Average Treatment Effects

In what follows, I will briefly review the LATE framework of Imbens and Angrist (1994) and Angrist *et al.* (1996), focusing on its extension to the case with additional covariates.

The population consists of four latent groups: always-takers, for whom $D(1) = 1$ and $D(0) = 1$; never-takers, for whom $D(1) = 0$ and $D(0) = 0$; compliers, for whom $D(1) = 1$ and $D(0) = 0$; and defiers, for whom $D(1) = 0$ and $D(0) = 1$. As demonstrated by Imbens and Angrist (1994), if we make several baseline assumptions on Z, rule out the existence of defiers, and assume that X is irrelevant, the estimand of interest, $\beta_{\text{IV}} = \beta_{\text{2SLS}}$, recovers the average treatment effect for compliers, usually referred to as the local average treatment effect (LATE).

Some of my results will allow for the existence of both compliers and defiers, and hence throughout this paper I follow Kolesár (2013) in defining LATE as

$$\tau_{\text{LATE}} = \mathrm{E}\left[Y(1) - Y(0) \mid D(1) \neq D(0)\right], \tag{7}$$

i.e. the average treatment effect for individuals whose treatment status is affected by the instrument. This group includes both compliers and defiers; it will be restricted to compliers whenever the existence of defiers is ruled out. It is also useful to write this unconditional LATE parameter as

$$\tau_{\text{LATE}} = \frac{\mathrm{E}\left[\pi(X) \cdot \tau(X)\right]}{\mathrm{E}\left[\pi(X)\right]}, \tag{8}$$

where

$$\tau(x) = \mathrm{E}\left[Y(1) - Y(0) \mid D(1) \neq D(0), X = x\right] \tag{9}$$

is the conditional LATE and

$$\pi(x) = \mathrm{P}\left[D(1) \neq D(0) \mid X = x\right] \tag{10}$$

is the conditional proportion of compliers and defiers. The following assumption, together with additional assumptions below, will be used to identify $\tau(x)$ and $\pi(x)$, and thereby also τ_{LATE}.

Assumption IV.

(i) (Conditional independence) $(Y(0,0), Y(0,1), Y(1,0), Y(1,1), D(0), D(1)) \perp Z \mid X$;

(ii) (Exclusion restriction) $\mathrm{P}[Y(1,D) = Y(0,D) \mid X] = 1$ for $D \in \{0,1\}$ a.s.;

(iii) (First stage) $0 < \mathrm{P}[Z = 1 \mid X] < 1$ and $\mathrm{P}[D(1) = 1 \mid X] \neq \mathrm{P}[D(0) = 1 \mid X]$ a.s.

Assumption IV(i) postulates that the instrument is "as good as randomly assigned" conditional on covariates. Assumption IV(ii) states that the instrument does not directly affect the outcome; its only effect on the outcome is through treatment status. Finally, Assumption IV(iii) requires that there is variation in encouragement to treatment as well as a distinct number of compliers and defiers at every value of covariates.

Assumption IV is not sufficient to identify $\tau(x)$ and $\pi(x)$. One possibility is to follow Imbens and Angrist (1994) in additionally ruling out the existence of defiers, as stated by Assumption UM.

Assumption UM (Unconditional monotonicity)**.** $\mathrm{P}[D(1) \geq D(0) \mid X] = 1$ a.s.

Assumption UM may often be too restrictive. In fact, Semenova (2020) provides evidence against this assumption in the National Job Corps Study. Kolesár (2013) and Semenova (2020) replace Assumption UM with a weaker assumption that postulates the existence of compliers but no defiers at some covariate values and the existence of defiers but no compliers elsewhere.

Assumption CM (Conditional monotonicity)**.** There exists a partition of covariate space such that $\mathrm{P}[D(1) \geq D(0) \mid X] = 1$ a.s. on one subset and $\mathrm{P}[D(1) \leq D(0) \mid X] = 1$ a.s. on its complement.

Assumption CM is weaker than Assumption UM, and yet, together with Assumption IV, it still allows us to identify $\tau(x)$ and $\pi(x)$. Before stating the relevant lemma, it is useful to define an auxiliary function

$$c(x) = \operatorname{sgn}\Big(\mathrm{P}[D(1) \geq D(0) \mid X = x] - \mathrm{P}[D(1) \leq D(0) \mid X = x]\Big), \tag{11}$$

where sgn(·) is the sign function. Clearly, $c(x)$ equals 1 if there are only compliers at $X = x$ and -1 if there are only defiers at $X = x$.

The following lemma summarizes identification of the conditional LATE parameter and the conditional proportion of individuals whose treatment status is affected by the instrument.

Lemma 2.1.

(i) *Under Assumptions IV and UM,* $\tau(x) = \beta(x)$ *and* $\pi(x) = \omega(x)$.

(ii) *Under Assumptions IV and CM,* $\tau(x) = \beta(x)$ *and* $\pi(x) = |\omega(x)| = c(x) \cdot \omega(x)$.

Lemma 2.1 consists of well-known results and straightforward extensions of these results, and as such it is stated without proof. The conditional Wald estimand identifies the conditional LATE parameter under both unconditional and conditional monotonicity. Under unconditional monotonicity, the conditional proportion of compliers is identified as the conditional first-stage slope coefficient. Under conditional monotonicity, the conditional proportion of compliers or defiers is identified as the absolute value of this coefficient; the coefficient is negative if and only if there are defiers but no compliers at a given value of covariates. Finally, it will be useful for what follows that $[\pi(x)]^2 = [\omega(x)]^2$ under either unconditional or conditional monotonicity.

2.3 Restricting Treatment Effect Heterogeneity

As established in the literature on marginal treatment effects, an alternative to monotonicity is to restrict the heterogeneity in treatment effects conditional on covariates (see, *e.g.*, Heckman and Vytlacil, 2007a,b; Mogstad and Torgovitsky, 2018). If we assume that the marginal treatment effect, *i.e.* the effect of treatment conditional on observables and unobservables, does not, in fact, depend on unobservables, then the conditional Wald estimand, $\beta(x)$, identifies the conditional average treatment effect, $\mathrm{E}[Y(1) - Y(0) \mid X = x]$. Under this assumption, we can identify and estimate the average treatment effect (ATE), since $\tau_{\mathrm{ATE}} = \mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[\mathrm{E}[Y(1) - Y(0) \mid X]]$.

It should be stressed, however, that this restriction on treatment effect heterogeneity is very strong. As discussed by Heckman and Vytlacil (2007a,b) and Mogstad and Torgovitsky (2018), it

implies that either $Y(1) - Y(0)$ is identical for all individuals with $X = x$ or these individuals do not select into treatment based on their unobserved returns from this treatment. In what follows, I will generally avoid making such assumptions, although I will explain how some of my results could be reinterpreted if I replaced monotonicity with this restriction on treatment effects.

3 Main Results

In this section I provide the main results of this paper. First, I revisit Angrist and Imbens (1995)'s representation of the 2SLS estimand and restate it in a form that is directly comparable to the unconditional LATE parameter in equation (8). Second, I demonstrate how this estimand changes when we retain Angrist and Imbens (1995)'s restriction that the model for covariates is saturated but no longer require that there is a separate first-stage coefficient on the instrument for every combination of covariate values. Third, I show that the same representation of the IV estimand is still appropriate when the model is not saturated but the instrument propensity score, *i.e.* the conditional probability that an individual is encouraged to get treated, is linear in covariates. In the latter two cases, under Assumption CM, the weights on some conditional LATEs may be negative, in which case the linear IV estimand is no longer interpretable as a causal effect. Finally, I demonstrate that the IV weights continue to be problematic for interpretation under Assumption UM. In this case, we cannot generally expect linear IV to recover the unconditional LATE parameter unless the groups with different values of the instrument are roughly equal sized.

3.1 Angrist and Imbens (1995), Revisited

Angrist and Imbens (1995) study a special case of the model in equation (1) where all covariates are binary and represent membership in disjoint groups or strata. In this framework, each of the original covariates needs to be discrete, in which case the population can be divided into K groups, where K corresponds to the number of possible combinations of values of these variables. (For example, with six binary variables, we have $K = 2^6 = 64$.) Let $G \in \{1, \ldots, K\}$ denote group

membership and $G_k = 1[G = k]$ denote the resulting group indicators. Angrist and Imbens (1995) consider a model where original covariates are replaced with these group indicators while reduced-form and first-stage regressions include a full set of interactions between these indicators and Z. Put another way, $X = (1, G_1, \ldots, G_{K-1})$ and $Z_C = (Z, ZG_1, \ldots, ZG_{K-1})$. As a result, we have a separate first-stage coefficient on Z for every value of X. The following lemma restates Angrist and Imbens (1995)'s and Kolesár (2013)'s interpretation of the 2SLS estimand in this context.

Lemma 3.1 (Angrist and Imbens, 1995; Kolesár, 2013)**.** *Under Assumptions IV and either UM or CM, and with* $X = (1, G_1, \ldots, G_{K-1})$ *and* $Z_C = (Z, ZG_1, \ldots, ZG_{K-1})$*,*

$$\beta_{2SLS} = \frac{\mathrm{E}\left[\sigma^2(X) \cdot \tau(X)\right]}{\mathrm{E}\left[\sigma^2(X)\right]},$$

where $\sigma^2(X) = \mathrm{E}[\mathrm{E}[D \mid X, Z] \cdot (\mathrm{E}[D \mid X, Z] - \mathrm{E}[D \mid X]) \mid X]$.

Lemma 3.1 establishes that the 2SLS estimand in the overidentified specification of Angrist and Imbens (1995) can be written as a convex combination of conditional LATEs, with weights equal to the conditional variance of the first stage. This result is due to Angrist and Imbens (1995) and has usually been interpreted as requiring Assumption UM (see, *e.g.*, Angrist and Pischke, 2009). Kolesár (2013) demonstrates that it also holds under Assumption CM.

A limitation of Lemma 3.1 is that it may not be immediately obvious how the 2SLS weights differ from the "desired" weights in equation (8). The following result, which is a straightforward implication of Lemma 3.1 but nonetheless appears to be novel, facilitates such a comparison.

Theorem 3.2. *Under Assumptions IV and either UM or CM, and with* $X = (1, G_1, \ldots, G_{K-1})$ *and* $Z_C = (Z, ZG_1, \ldots, ZG_{K-1})$*,*

$$\beta_{2SLS} = \frac{\mathrm{E}\left[[\pi(X)]^2 \cdot \mathrm{Var}[Z \mid X] \cdot \tau(X)\right]}{\mathrm{E}\left[[\pi(X)]^2 \cdot \mathrm{Var}[Z \mid X]\right]}.$$

Proof. Lemma 3.1 states that $\beta_{2SLS} = \frac{\mathrm{E}[\sigma^2(X) \cdot \tau(X)]}{\mathrm{E}[\sigma^2(X)]}$. It remains to show that $\sigma^2(X) = [\pi(X)]^2 \cdot \mathrm{Var}[Z \mid X]$. Indeed, it follows from the definition of $\sigma^2(X)$, equation (4), and iterated expectations

that $\sigma^2(X) = [\omega(X)]^2 \cdot \text{Var}[Z \mid X]$. Then, it follows from Lemma 2.1 that $\sigma^2(X) = [\pi(X)]^2 \cdot \text{Var}[Z \mid X]$ because $[\omega(X)]^2 = [\pi(X)]^2$ under Assumptions IV and either UM or CM. □

Theorem 3.2 shows that the 2SLS estimand in Angrist and Imbens (1995)'s specification is a convex combination of conditional LATEs, with weights equal to the product of the squared conditional proportion of compliers or defiers and the conditional variance of Z. Since the "desired" weights, as shown in equation (8), consist only of the conditional proportion of compliers or defiers, Angrist and Imbens (1995)'s specification overweights the effects in groups with strong first stages (*i.e.* many individuals affected by the instrument) and with large variances of Z.

Remark 3.1. A major limitation of Lemma 3.1 and Theorem 3.2 is that empirical applications of IV methods rarely consider fully heterogeneous first stages and saturated specifications with discrete covariates. For example, in a survey of recent papers with multiple instruments, only 13% of applications use covariate interactions with an original instrument (Mogstad, Torgovitsky, and Walters, 2020). Specifications using many overidentifying restrictions appear to have been more common in earlier work using IV methods (*e.g.*, Angrist, 1990; Angrist and Krueger, 1991) but have effectively disappeared from empirical economics out of concern for weak instruments.[2]

Remark 3.2. If we replace either of the monotonicity assumptions with an appropriate restriction on treatment effect heterogeneity, as discussed in Section 2.3, the 2SLS estimand in Angrist and Imbens (1995)'s specification will correspond to a convex combination of conditional ATEs, with weights equal to the product of the squared conditional first-stage slope coefficient and the conditional variance of Z. Since the (unconditional) average treatment effect, $\tau_{\text{ATE}} = \text{E}[Y(1) - Y(0)]$, is an unweighted average of conditional ATEs, this weighting is generally undesirable.

3.2 Results for Just Identified Models

Remark 3.1 suggests that Theorem 3.2 is not necessarily useful for interpreting actual empirical studies because modern applications of IV methods avoid using many overidentifying restrictions.

[2]Indeed, Bound, Jaeger, and Baker (1995) write that their results "indicate that *the common practice* of adding interaction terms as excluded instruments may exacerbate the problem [of finite-sample biases]" (emphasis mine).

A similar point is made by Angrist and Pischke (2009, p. 178), who write that "[i]n practice, we may not want to work with a model with a first-stage parameter for each value of the covariates... It seems reasonable to imagine that models with fewer parameters, say a restricted first stage imposing a constant [effect of Z on D], nevertheless approximate some kind of covariate-averaged LATE. This turns out to be true, but the argument [due to Abadie (2003)] is surprisingly indirect." In what follows, I will show that this claim is *false* whenever the monotonicity assumption is valid only conditional on covariates. The claim is true under unconditional monotonicity, which I will be able to demonstrate directly. I will return to Abadie (2003)'s indirect argument later on.

Let us first consider the case of conditional monotonicity. The following result clarifies the lack of causal interpretability of the linear IV estimand in this context. For now, I assume that the model for covariates remains saturated.

Theorem 3.3. *Under Assumptions IV and CM, and with* $X = (1, G_1, \ldots, G_{K-1})$*,*

$$\beta_{\text{IV}} = \frac{\mathrm{E}\left[c(X) \cdot \pi(X) \cdot \mathrm{Var}\left[Z \mid X\right] \cdot \tau(X)\right]}{\mathrm{E}\left[c(X) \cdot \pi(X) \cdot \mathrm{Var}\left[Z \mid X\right]\right]}.$$

Proof. See Appendix A. □

Theorem 3.3 provides a new representation of the IV estimand in the standard specification, *i.e.* one that, perhaps incorrectly, restricts the effects of the instrument in the reduced-form and first-stage regressions to be homogeneous. For ease of comparison, the only difference between the specifications in Theorems 3.2 and 3.3 is in their treatment of reduced-form and first-stage heterogeneity. Unlike in Angrist and Imbens (1995)'s specification, the estimand in the standard specification is not necessarily a convex combination of conditional LATEs. This is because $c(X)$ takes the value of -1 for every value of covariates where there exist defiers but no compliers, and hence the weights attached to all corresponding LATEs are negative as well. It follows that, when IV is applied in the usual way, the estimand may no longer be interpretable as a causal effect for any subpopulation.

The following result demonstrates that this problem disappears when we impose the unconditional version of monotonicity.

Corollary 3.4. *Under Assumptions IV and UM, and with* $X = (1, G_1, \ldots, G_{K-1})$*,*

$$\beta_{\text{IV}} = \frac{\text{E}\left[\pi(X) \cdot \text{Var}\left[Z \mid X\right] \cdot \tau(X)\right]}{\text{E}\left[\pi(X) \cdot \text{Var}\left[Z \mid X\right]\right]}.$$

Proof. Note that Assumption UM is a special case of Assumption CM where the existence of compliers but no defiers is postulated at all covariate values and the existence of defiers but no compliers everywhere else (*i.e.* on an empty set). Thus, it follows from Theorem 3.3 that, under Assumptions IV and UM, $\beta_{\text{IV}} = \frac{\text{E}[c(X)\cdot\pi(X)\cdot\text{Var}[Z|X]\cdot\tau(X)]}{\text{E}[c(X)\cdot\pi(X)\cdot\text{Var}[Z|X]]}$ and $c(X) = 1$ a.s. □

Corollary 3.4 provides a direct argument for Angrist and Pischke (2009)'s assertion that the standard specification of IV recovers a convex combination of conditional LATEs. As noted previously, however, this statement is no longer true under conditional monotonicity. If unconditional monotonicity holds, then the weights in Corollary 3.4 are more desirable than those in Angrist and Imbens (1995)'s specification. Indeed, a comparison of Corollary 3.4 and equation (8) shows that the standard specification, like Angrist and Imbens (1995)'s specification, overweights the effects in groups with large variances of Z but not, unlike the latter, in groups with strong first stages.

Remark 3.3. Much of the literature on weak and many instruments replicates an overidentified specification from Angrist and Krueger (1991) and compares the estimates with those from a model with a small number of restrictions.[3] The original study considers a model with 180 instruments which have been created by interacting selected covariates with quarter-of-birth indicators (*i.e.* the original instruments). Bound *et al.* (1995) argue that this specification suffers from finite-sample biases. A comparison of Theorems 3.2 and 3.3 implies that the difference between a similar specification and the standard specification can also be explained by the underlying weights.[4]

Remark 3.4. Bond, White, and Walker (2007) discuss the interpretation of an overidentified and a just identified specification in randomized experiments with noncompliance and unconditional

[3] See, *e.g.*, Bound *et al.* (1995), Staiger and Stock (1997), Donald and Newey (2001), Cruz and Moreira (2005), and Hansen, Hausman, and Newey (2008).

[4] Because the overidentified specification from Angrist and Krueger (1991) does not exactly fit into the framework of Angrist and Imbens (1995) and Theorem 3.2, I do not replicate it in this paper.

monotonicity. In this case, the standard specification of IV recovers the unconditional LATE parameter but the overidentified specification does not. This is a special case of the difference between Theorem 3.2 and Corollary 3.4 where $\text{Var}[Z \mid X]$ is constant. However, Theorem 3.3 makes it clear that under conditional monotonicity the standard specification no longer recovers the unconditional LATE parameter or even a convex combination of conditional LATEs.

3.3 Results for Nonsaturated Models

The theoretical results have so far relied on an impractical restriction that all covariates are discrete and the model for covariates is saturated. In what follows, I will show that the representation of the IV estimand in Theorem 3.3 and Corollary 3.4 remains unchanged when we replace this restriction with the assumption that the instrument propensity score, defined as

$$e(x) = \text{E}[Z \mid X = x], \tag{12}$$

i.e. the conditional probability that an individual is encouraged to get treated, is linear in X.

Assumption PS (Instrument propensity score)**.** $e(X) = X\alpha$.

Assumption PS holds automatically when Z is randomized, and also when all covariates are discrete and the model for covariates is saturated. It may also provide a good approximation to $e(X)$ in other situations, especially when X includes powers and cross-products of original covariates. This assumption has been used by Abadie (2003), Kolesár (2013), Lochner and Moretti (2015), and Evdokimov and Kolesár (2019), among others. It allows us to establish the following result.

Theorem 3.5. *Under Assumptions IV, CM, and PS,*

$$\beta_{\text{IV}} = \frac{\text{E}[c(X) \cdot \pi(X) \cdot \text{Var}[Z \mid X] \cdot \tau(X)]}{\text{E}[c(X) \cdot \pi(X) \cdot \text{Var}[Z \mid X]]}.$$

Proof. See Appendix A. □

Theorem 3.5 provides a reaffirmation of Theorem 3.3 when the model for covariates may not be saturated. As before, the IV estimand is not necessarily a convex combination of conditional LATEs. It is possible that this parameter may turn out to be negative (positive) even if treatment effects are positive (negative) for everyone in the population. The following result reiterates the point of Corollary 3.4 that this problem disappears when we impose Assumption UM.

Corollary 3.6. *Under Assumptions IV, UM, and PS,*

$$\beta_{\text{IV}} = \frac{\text{E}\left[\pi(X) \cdot \text{Var}\left[Z \mid X\right] \cdot \tau(X)\right]}{\text{E}\left[\pi(X) \cdot \text{Var}\left[Z \mid X\right]\right]}.$$

The proof of Corollary 3.6 is analogous to that of Corollary 3.4 and is therefore omitted. This result makes it clear that under unconditional monotonicity the standard specification of IV recovers a convex combination of conditional LATEs, with weights that differ from the "desired" weights in equation (8) only through their dependence on $\text{Var}\left[Z \mid X\right]$.

In what follows, I will compare my results in Theorem 3.5 and Corollary 3.6 with other papers that rely on Assumption PS. I will also discuss the interpretation of the IV estimand when either of the monotonicity assumptions is replaced with a restriction on treatment effect heterogeneity.

Remark 3.5. Abadie (2003) shows that, under Assumptions IV, UM, and PS, the IV estimand is equivalent to the coefficient on D in the linear projection of Y on D and X among compliers.[5] In other words, IV is analogous to ordinary least squares (OLS), with the exception of its focus on this latent group. Corollary 3.6 provides another argument that "IV is like OLS." Indeed, as shown by Angrist (1998), the difference between the OLS estimand and ATE is due to the dependence of the OLS weights on the conditional variance of D. Similarly, as shown in Corollary 3.6, the difference between the IV estimand and LATE is due to the dependence of the IV weights on the conditional variance of Z. The analogy between OLS and IV is potentially problematic for IV given the results on OLS in my earlier work (Słoczyński, 2020). I will return to this point later on.

[5]To be precise, Abadie (2003)'s formulation of what I refer to as Assumption IV(iii) is slightly different but this is not consequential in the present context.

Remark 3.6. Kolesár (2013) provides a general representation of linear IV and 2SLS estimands under conditional monotonicity, and concludes that they can be written as a convex combination of conditional LATEs "under some mild assumptions about the first stage." This conclusion seems at odds with Theorem 3.5, which shows that under conditional monotonicity some of the IV weights may be negative. However, Kolesár (2013)'s condition for positive weights is that the first stage postulated by the researcher is monotone in the true first stage. Under conditional monotonicity, and when some defiers do indeed exist, this requirement is easily seen to never be satisfied in the standard specification of IV, which limits the first-stage effects of Z on D to be homogeneous.

Remark 3.7. Kolesár (2013) and Evdokimov and Kolesár (2019) also consider a special case of this problem where the reduced-form and first-stage regressions are correctly specified. How does this assumption affect the interpretation of the IV estimand? Recall that the standard specification of IV corresponds to reduced-form and first-stage regressions that limit the effects of Z on Y and D to be homogeneous. This is consistent with unconditional monotonicity, and with $\pi(X)$ and $\tau(X)$ that do not depend on X. Thus, following Corollary 3.6, $\beta_{\text{IV}} = \frac{\text{E}[\text{Var}[Z|X]\cdot\tau(X)]}{\text{E}[\text{Var}[Z|X]]} = \tau_{\text{LATE}}$ because $\tau(X) = \tau_{\text{LATE}}$ for all covariate values. Still, these homogeneity assumptions are rather implausible.

Remark 3.8. Lochner and Moretti (2015) show that under Assumption PS the unconditional IV estimand is equivalent to a weighted average of conditional IV estimands, with weights equal to the conditional covariance between the treatment and the instrument.[6] This equivalence is also implicit in Theorem 3.5 and Corollary 3.6.

Remark 3.9. Hull (2018) recommends that researchers consider an alternative estimation method, namely the IV regression of Y on 1 and D, with $Z - e(X)$ as the instrument for D. Under unconditional monotonicity, as also shown by Hull (2018), this estimator converges to a convex combination of conditional LATEs, with weights equal to the product of the conditional proportion of compliers and the conditional variance of Z. However, as shown in Corollary 3.6, the standard specification of IV recovers *exactly* the same parameter. Note that, even though Hull (2018)

[6]Related results are also discussed by Kling (2001) and Ishimaru (2019). Note, however, that none of these papers, including Lochner and Moretti (2015), uses the LATE framework to interpret the IV estimand.

does not invoke Assumption PS, it can be shown that Hull (2018)'s estimator is also numerically identical to standard IV, as long as the linear probability model and OLS are used to estimate $e(X)$.

Remark 3.10. If we replace Assumptions CM and UM with a restriction on treatment effect heterogeneity, as discussed in Section 2.3 and Remark 3.2, the IV estimand will correspond to a weighted average of conditional ATEs, with weights equal to the product of the conditional first-stage slope coefficient and the conditional variance of Z. Unlike in Angrist and Imbens (1995)'s specification, some of these weights may be negative. To ensure positive weights, it is sufficient to additionally impose a weaker version of unconditional monotonicity, where we only require that all conditional first-stage slope coefficients have the same sign or, equivalently, that there are more compliers than defiers (or more defiers than compliers) at all covariate values.

3.4 Further Results for Models with Unconditional Monotonicity

The theoretical analysis so far makes it clear that the IV estimand in the standard specification is not necessarily a convex combination of conditional LATEs under conditional monotonicity. If monotonicity holds unconditionally, however, the IV weights are guaranteed to be positive. In what follows, I will argue that, even in this optimistic scenario, we should not interpret the IV estimand as if it was somehow guaranteed to (approximately) correspond to the unconditional LATE parameter. This is because the IV weights are still systematically different from the "desired" weights in equation (8). Indeed, I will demonstrate that the IV estimand may be substantially different from the unconditional LATE parameter unless the groups that are encouraged and not encouraged to get treated are roughly equal sized.

The starting point is to introduce an additional parameter, namely the local average treatment effect on the treated (LATT), previously discussed by Frölich and Lechner (2010), Hong and Nekipelov (2010), and Donald, Hsu, and Lieli (2014). We can define LATT as follows:

$$\tau_{\text{LATT}} = \mathrm{E}\left[Y(1) - Y(0) \mid D(1) \neq D(0), D = 1\right]. \tag{13}$$

It is also useful to define the local average treatment effect on the untreated (LATU) as

$$\tau_{\text{LATU}} = \mathrm{E}\left[Y(1) - Y(0) \mid D(1) \neq D(0), D = 0\right]. \tag{14}$$

Clearly, the unconditional LATE parameter is a convex combination of LATT and LATU; that is,

$$\tau_{\text{LATE}} \;=\; \mathrm{P}\left[D = 1 \mid D(1) \neq D(0)\right] \cdot \tau_{\text{LATT}} \;+\; \mathrm{P}\left[D = 0 \mid D(1) \neq D(0)\right] \cdot \tau_{\text{LATU}}. \tag{15}$$

Under Assumptions IV and UM, we can also represent LATT and LATU as

$$\begin{aligned}
\tau_{\text{LATT}} &= \mathrm{E}\left[Y(1) - Y(0) \mid D(1) > D(0), D = 1\right] \\
&= \mathrm{E}\left[Y(1) - Y(0) \mid D(1) > D(0), Z = 1\right] \\
&= \frac{\mathrm{E}\left[\pi(X) \cdot \tau(X) \mid Z = 1\right]}{\mathrm{E}\left[\pi(X) \mid Z = 1\right]} \\
&= \frac{\mathrm{E}\left[e(X) \cdot \pi(X) \cdot \tau(X)\right]}{\mathrm{E}\left[e(X) \cdot \pi(X)\right]}
\end{aligned} \tag{16}$$

and

$$\begin{aligned}
\tau_{\text{LATU}} &= \mathrm{E}\left[Y(1) - Y(0) \mid D(1) > D(0), D = 0\right] \\
&= \mathrm{E}\left[Y(1) - Y(0) \mid D(1) > D(0), Z = 0\right] \\
&= \frac{\mathrm{E}\left[\pi(X) \cdot \tau(X) \mid Z = 0\right]}{\mathrm{E}\left[\pi(X) \mid Z = 0\right]} \\
&= \frac{\mathrm{E}\left[(1 - e(X)) \cdot \pi(X) \cdot \tau(X)\right]}{\mathrm{E}\left[(1 - e(X)) \cdot \pi(X)\right]}.
\end{aligned} \tag{17}$$

The first equality in equations (16) and (17) follows from Assumption UM. The second equality uses the fact that all treated compliers are encouraged to get treated and all untreated compliers are not (call this "DZ equivalence"). The third and fourth equalities follow from Assumption IV, iterated expectations, and a little algebra. We can also use Assumption UM, DZ equivalence, and

Bayes' rule to rewrite equation (15) as

$$\tau_{\text{LATE}} = \frac{\theta \cdot \pi_1}{\theta \cdot \pi_1 + (1-\theta) \cdot \pi_0} \cdot \tau_{\text{LATT}} + \frac{(1-\theta) \cdot \pi_0}{\theta \cdot \pi_1 + (1-\theta) \cdot \pi_0} \cdot \tau_{\text{LATU}}, \tag{18}$$

where

$$\theta = \text{P}[Z = 1] \tag{19}$$

is the proportion of the population that is encouraged to get treated and

$$\pi_z = \text{P}[D(1) > D(0) \mid Z = z] \tag{20}$$

is the proportion of compliers in the subpopulation with $Z = z$.

In what follows, I will develop two arguments to show that the IV weights in Corollary 3.6 continue to be problematic for interpretation. The starting point for my first argument is to observe that $\text{Var}[Z \mid X] = e(X) \cdot (1 - e(X))$. Then, note that $\text{Var}[Z \mid X] \approx e(X)$ if $e(X)$ is close to zero and, similarly, $\text{Var}[Z \mid X] \approx 1 - e(X)$ if $e(X)$ is close to one. These approximations are important because the only difference between the IV estimand in Corollary 3.6 and the parameters in equations (16) and (17) is in their respective use of $\text{Var}[Z \mid X]$, $e(X)$, and $1 - e(X)$ to reweight the product of $\pi(X)$ and $\tau(X)$. This observation implies that, when $e(X)$ is close to zero or one for all covariate values, which also means that θ is close to zero or one, the IV estimand in Corollary 3.6 is similar to LATT or LATU, respectively. Perhaps surprisingly, when θ is close to zero (one) or, in other words, almost no (almost all) individuals are encouraged to get treated, the IV estimand is similar to the local average treatment effect on the treated (untreated). This is the opposite of what we want if our goal is to recover the unconditional LATE parameter, as represented in equation (18).

The informal argument above parallels a remark of Humphreys (2009) about the interpretation of the OLS estimand under unconfoundedness. My second argument formalizes this discussion by demonstrating that under an additional assumption the IV estimand can be written as a convex combination of LATT and LATU, with weights that, compared with equation (18), are related to

θ in the opposite direction. Namely, the greater the value of θ, the greater is the contribution of LATT to the unconditional LATE parameter and yet the smaller is the IV weight on LATT. The following assumption will be useful for establishing this result.

Assumption LN.

(i) (Reduced form) $\mathrm{E}[Y \mid X, Z] = \delta_1 + \delta_2 Z + \delta_3 \cdot e(X) + \delta_4 Z \cdot e(X)$;

(ii) (First stage) $\mathrm{E}[D \mid X, Z] = \eta_1 + \eta_2 Z + \eta_3 \cdot e(X) + \eta_4 Z \cdot e(X)$.

Assumption LN postulates that the true reduced-form and first-stage regressions are linear in $e(X)$ conditional on Z. This assumption is fairly strong, although a similar restriction on potential outcomes under unconfoundedness, *i.e.* that they are linear in the propensity score, has been used by Rosenbaum and Rubin (1983) and Słoczyński (2020). The following result confirms that the IV estimand "reverses" the role of θ in the implicit weights on LATT and LATU.

Theorem 3.7. *Under Assumptions IV, UM, PS, and LN,*

$$\beta_{\mathrm{IV}} \;=\; w_{\mathrm{LATT}} \cdot \tau_{\mathrm{LATT}} \;+\; w_{\mathrm{LATU}} \cdot \tau_{\mathrm{LATU}},$$

where $w_{\mathrm{LATT}} = \frac{(1-\theta)\cdot \mathrm{Var}[e(X)|Z=0]\cdot\pi_1}{\theta\cdot \mathrm{Var}[e(X)|Z=1]\cdot\pi_0 \,+\, (1-\theta)\cdot \mathrm{Var}[e(X)|Z=0]\cdot\pi_1}$ *and* $w_{\mathrm{LATU}} = \frac{\theta\cdot \mathrm{Var}[e(X)|Z=1]\cdot\pi_0}{\theta\cdot \mathrm{Var}[e(X)|Z=1]\cdot\pi_0 \,+\, (1-\theta)\cdot \mathrm{Var}[e(X)|Z=0]\cdot\pi_1}$.

Proof. See Appendix A. □

Theorem 3.7 provides an alternative representation of the IV estimand under unconditional monotonicity. Unlike in Corollary 3.6, it now follows immediately that the IV weights are potentially very problematic. The first thing to note is that the weights are always positive and sum to one. Then, however, it turns out that the weight on LATT is increasing in $\frac{\pi_1}{\pi_0}$, which is desirable; decreasing in $\frac{\mathrm{Var}[e(X)|Z=1]}{\mathrm{Var}[e(X)|Z=0]}$, which I largely ignore for simplicity; and decreasing in θ, which is potentially alarming. Because $w_{\mathrm{LATU}} = 1 - w_{\mathrm{LATT}}$, the weight on LATU always changes in the opposite direction. This result, and some of the subsequent discussion, parallels my earlier work on the

interpretation of the OLS estimand under unconfoundedness (Słoczyński, 2020), which demonstrates that this parameter can be written as a convex combination of the average treatment effects on the treated (ATT) and untreated (ATU).

An implication of Theorem 3.7 is that we can express the difference between the IV estimand and the unconditional LATE parameter as a product of a particular measure of heterogeneity in conditional LATEs, *i.e.* the difference between LATT and LATU, and an additional parameter that is equal to the difference between the actual and the "desired" weight on LATT.

Corollary 3.8. *Under Assumptions IV, UM, PS, and LN,*

$$\beta_{\text{IV}} - \tau_{\text{LATE}} \;=\; \lambda \cdot (\tau_{\text{LATT}} - \tau_{\text{LATU}}),$$

where $\lambda = \frac{(1-\theta)\cdot \text{Var}[e(X)|Z=0]\cdot \pi_1}{\theta\cdot \text{Var}[e(X)|Z=1]\cdot \pi_0 \,+\, (1-\theta)\cdot \text{Var}[e(X)|Z=0]\cdot \pi_1} - \frac{\theta\cdot \pi_1}{\theta\cdot \pi_1 \,+\, (1-\theta)\cdot \pi_0}$.

The proof of Corollary 3.8 follows from simple algebra and is omitted. This result specifies the conditions under which the IV estimand recovers the unconditional LATE parameter. One possibility is that the local average treatment effects on the treated (LATT) and untreated (LATU) are identical. Another possibility is that the IV weights on LATT and LATU correspond to the "desired" weights in equation (18), which would imply that $\lambda = 0$. The following restriction, which requires that the conditional variance of $e(X)$ is the same among the individuals that are encouraged and not encouraged to get treated, will allow us to simplify the formula for λ.

Assumption EV (Equality of variances)**.** $\text{Var}\left[e(X) \mid Z = 1\right] = \text{Var}\left[e(X) \mid Z = 0\right]$.

Indeed, under Assumption EV, simple algebra shows that $\lambda = \frac{(1-2\theta)\,\cdot\, \pi_0\pi_1}{(\theta\cdot\pi_0 \,+\, (1-\theta)\cdot\pi_1)\,\cdot\, (\theta\cdot\pi_1 \,+\, (1-\theta)\cdot\pi_0)}$. Clearly, the only case where the IV weights overlap with the "desired" weights, or $\lambda = 0$, occurs when the groups that are encouraged and not encouraged to get treated are equal sized, $\theta = 0.5$. The following result makes it clear that, under Assumption EV, the IV estimand recovers the unconditional LATE parameter if and only if $\theta = 0.5$ or LATT and LATU are identical.

Corollary 3.9. *Under Assumptions IV, UM, PS, LN, and EV,*

$$\beta_{\text{IV}} = \tau_{\text{LATE}} \qquad \textit{if and only if} \qquad \tau_{\text{LATT}} = \tau_{\text{LATU}} \;\; \textit{or} \;\; \theta = 0.5.$$

Proof. See Appendix A. □

Corollary 3.9 shows that under certain assumptions the IV estimand can be interpreted as the unconditional LATE parameter only when either of two restrictive conditions is satisfied, $\theta = 0.5$ or $\tau_{\text{LATT}} = \tau_{\text{LATU}}$. Even if one or more of the assumptions in Corollary 3.9 are not exactly true, they might be approximately true, in which case the value of θ might provide a useful rule of thumb for the interpretation of the IV estimand. For example, when the groups with different values of the instrument are roughly equal sized, or $\theta \approx 0.5$, we may be willing to interpret the IV estimand as LATE, but not otherwise. See also Słoczyński (2020) for a related discussion of OLS.

4 Conclusion

In this paper I study the interpretation of linear IV and 2SLS estimands when both the endogenous treatment and the instrument are binary, and when additional covariates are required for identification. I follow the LATE framework of Imbens and Angrist (1994) and Angrist *et al.* (1996), and conclude that the common practice of interpreting linear IV and 2SLS estimands as a convex combination of conditional LATEs, or even as an "overall" LATE, is substantially more problematic than previously thought. For example, Kolesár (2013) concludes that the weights on all conditional LATEs are guaranteed to be positive "under some mild assumptions about the first stage," even when the monotonicity assumption of Imbens and Angrist (1994) is valid only conditional on covariates. In this paper I show, among other things, that these "mild assumptions" are virtually guaranteed not to be satisfied in the usual application of instrumental variables that limits the effects of the instrument in the reduced-form and first-stage regressions to be homogeneous. Consequently, under conditional monotonicity, the linear IV estimand may no longer be interpretable

as a causal effect; this parameter may turn out to be negative (positive) even if treatment effects are positive (negative) for everyone in the population.

There are several lessons to be learned from my theoretical results. Empirical researchers with a preference for linear IV/2SLS may choose either of two paths to continue interpreting their estimands as a convex combination of conditional LATEs. One is to strengthen Kolesár (2013)'s assumption of conditional monotonicity and require that there are no defiers at any combination of covariate values. Another is to account for possible heterogeneity in the reduced-form and first-stage regressions, as in the overidentified specification of Angrist and Imbens (1995). Unfortunately, neither of these solutions guarantees that the resulting estimand will necessarily be similar to the "overall" (unconditional) LATE. If this is a concern, and I believe it should be, then my results also suggest that we may be able to claim similarity between the IV estimand and the unconditional LATE parameter when the groups that are encouraged and not encouraged to get treated (*i.e.* with different values of the instrument) are roughly equal sized.

If none of these solutions is appealing in a specific empirical context, it may be reasonable to give up on linear IV and 2SLS altogether. There are many alternative estimators of the unconditional LATE parameter that are available under unconditional monotonicity (*e.g.*, Abadie, 2003; Frölich, 2007). To the best of my knowledge, there are currently no estimators of LATE that are guaranteed to be consistent under conditional monotonicity. It is straightforward to construct such an estimator when all covariates are discrete. An extension to the case with continuous covariates is left for future research.

Appendix A Proofs

Proof of Theorem 3.3. Let R and T be generic notation for two random variables, where T is binary and R is arbitrarily discrete or continuous. The following lemma, due to Angrist (1998), will be useful for what follows.

Lemma A.1 (Angrist, 1998)**.** *Suppose that $X = (1, G_1, \ldots, G_{K-1})$. Then, ξ, the coefficient on T in the regression of R on T and X can be written as*

$$\xi = \frac{\mathrm{E}\left[\mathrm{Var}\left[T \mid X\right] \cdot \xi(X)\right]}{\mathrm{E}\left[\mathrm{Var}\left[T \mid X\right]\right]},$$

where $\xi(X) = \mathrm{E}\left[R \mid T = 1, X\right] - \mathrm{E}\left[R \mid T = 0, X\right]$.

Recall that β_{IV} is equal to the ratio of the reduced-form and first-stage coefficients on Z. It follows that we can apply Lemma A.1 separately to these two coefficients, and thereby obtain the following expression for the estimand of interest:

$$\beta_{\mathrm{IV}} = \frac{\frac{\mathrm{E}\left[\mathrm{Var}[Z|X] \cdot \phi(X)\right]}{\mathrm{E}[\mathrm{Var}[Z|X]]}}{\frac{\mathrm{E}[\mathrm{Var}[Z|X] \cdot \omega(X)]}{\mathrm{E}[\mathrm{Var}[Z|X]]}}, \tag{21}$$

where

$$\phi(X) = \mathrm{E}\left[Y \mid Z = 1, X\right] - \mathrm{E}\left[Y \mid Z = 0, X\right] \tag{22}$$

is the conditional reduced-form slope coefficient and $\omega(X)$ is as defined in equation (5). Upon rearrangement, we obtain

$$\begin{aligned} \beta_{\mathrm{IV}} &= \frac{\mathrm{E}\left[\mathrm{Var}\left[Z \mid X\right] \cdot \phi(X)\right]}{\mathrm{E}\left[\mathrm{Var}\left[Z \mid X\right] \cdot \omega(X)\right]} \\ &= \frac{\mathrm{E}\left[\mathrm{Var}\left[Z \mid X\right] \cdot \omega(X) \cdot \beta(X)\right]}{\mathrm{E}\left[\mathrm{Var}\left[Z \mid X\right] \cdot \omega(X)\right]}, \end{aligned} \tag{23}$$

where the second equality uses the definition of $\beta(X)$ in equation (6). Finally, we know from Lemma 2.1 that $\beta(X) = \tau(X)$ and $\omega(X) = c(X) \cdot \pi(X)$ under Assumptions IV and CM. This

completes the proof because the estimand of interest can now be written as

$$\beta_{\text{IV}} = \frac{\mathrm{E}\left[c(X) \cdot \pi(X) \cdot \mathrm{Var}\left[Z \mid X\right] \cdot \tau(X)\right]}{\mathrm{E}\left[c(X) \cdot \pi(X) \cdot \mathrm{Var}\left[Z \mid X\right]\right]}. \tag{24}$$

Proof of Theorem 3.5. Aronow and Samii (2016) demonstrate that the result in Lemma A.1 follows also in the case when the restriction that $X = (1, G_1, \ldots, G_{K-1})$ is relaxed but instead we assume that $\mathrm{E}\left[T \mid X\right]$ is linear in X. In our case, where Z plays the role of T in both the reduced-form and first-stage regressions, this restriction is equivalent to Assumption PS. Otherwise the proof is identical to that of Theorem 3.3.

Proof of Theorem 3.7. Let us continue using the same notation as in the proofs above. If $\mathrm{L}\left[\cdot \mid \cdot\right]$ denotes the linear projection, let $p(X)$ denote the best linear approximation to the "propensity score" for T; that is,

$$p(X) = \mathrm{L}\left[T \mid X\right] = X\rho, \tag{25}$$

with X being again completely general and not necessarily consisting only of group indicators. We also need two linear projections of R on 1 and $p(X)$, separately for $T = 1$ and $T = 0$; that is,

$$\mathrm{L}\left[R \mid 1, p(X), T = t\right] = \iota_t + \zeta_t \cdot p(X). \tag{26}$$

The following lemma, due to Słoczyński (2020), will be useful for what follows.

Lemma A.2 (Słoczyński, 2020)**.** *The coefficient on T in the regression of R on T and X, denoted by ξ, can be written as*

$$\begin{aligned} \xi &= w_1 \cdot \left((\iota_1 - \iota_0) + (\zeta_1 - \zeta_0) \cdot \mathrm{E}\left[p(X) \mid T = 1\right] \right) \\ &+ w_0 \cdot \left((\iota_1 - \iota_0) + (\zeta_1 - \zeta_0) \cdot \mathrm{E}\left[p(X) \mid T = 0\right] \right), \end{aligned}$$

where $w_1 = \frac{\mathrm{P}[T=0] \cdot \mathrm{Var}\left[p(X) \mid T=0\right]}{\mathrm{P}[T=1] \cdot \mathrm{Var}\left[p(X) \mid T=1\right] + \mathrm{P}[T=0] \cdot \mathrm{Var}\left[p(X) \mid T=0\right]}$ *and* $w_0 = \frac{\mathrm{P}[T=1] \cdot \mathrm{Var}\left[p(X) \mid T=1\right]}{\mathrm{P}[T=1] \cdot \mathrm{Var}\left[p(X) \mid T=1\right] + \mathrm{P}[T=0] \cdot \mathrm{Var}\left[p(X) \mid T=0\right]}$.

Again, we can use the fact that β_{IV} is equal to the ratio of the reduced-form and first-stage coefficients on Z, and apply Lemma A.2 separately to these coefficients. Thus, Y will play the role of R in the reduced-form regression, D will play the role of R in the first-stage regression, and Z will play the role of T in both regressions. Additionally, under Assumption PS, equation (25) corresponds to the true instrument propensity score and, under Assumption LN, equation (26) represents the true reduced-form and first-stage regressions. It follows from Lemma A.2 that under these assumptions the reduced-form and first-stage coefficients on Z are equal to a convex combination of the average causal effects of Z on Y and D in the subpopulations with $Z = 1$ and $Z = 0$, with weights equal to $w_1^* = \frac{(1-\theta)\cdot\text{Var}[e(X)|Z=0]}{\theta\cdot\text{Var}[e(X)|Z=1] + (1-\theta)\cdot\text{Var}[e(X)|Z=0]}$ and $w_0^* = \frac{\theta\cdot\text{Var}[e(X)|Z=1]}{\theta\cdot\text{Var}[e(X)|Z=1] + (1-\theta)\cdot\text{Var}[e(X)|Z=0]}$, respectively. Indeed,

$$\begin{aligned}
\beta_{\text{IV}} &= \frac{w_1^* \cdot \text{E}\left[Y(D(1)) - Y(D(0)) \mid Z = 1\right] \;+\; w_0^* \cdot \text{E}\left[Y(D(1)) - Y(D(0)) \mid Z = 0\right]}{w_1^* \cdot \pi_1 \;+\; w_0^* \cdot \pi_0} \\
&= \frac{w_1^* \cdot \pi_1 \cdot \tau_{\text{LATT}} \;+\; w_0^* \cdot \pi_0 \cdot \tau_{\text{LATU}}}{w_1^* \cdot \pi_1 \;+\; w_0^* \cdot \pi_0} \\
&= \frac{(1-\theta)\cdot\text{Var}\left[e(X) \mid Z = 0\right]\cdot\pi_1\cdot\tau_{\text{LATT}} \;+\; \theta\cdot\text{Var}\left[e(X) \mid Z = 1\right]\cdot\pi_0\cdot\tau_{\text{LATU}}}{(1-\theta)\cdot\text{Var}\left[e(X) \mid Z = 0\right]\cdot\pi_1 \;+\; \theta\cdot\text{Var}\left[e(X) \mid Z = 1\right]\cdot\pi_0} \\
&= w_{\text{LATT}} \cdot \tau_{\text{LATT}} \;+\; w_{\text{LATU}} \cdot \tau_{\text{LATU}}, \qquad (27)
\end{aligned}$$

where the second equality uses the fact that $\tau_{\text{LATT}} = \frac{\text{E}[Y(D(1))-Y(D(0))|Z=1]}{\text{E}[D(1)-D(0)|Z=1]}$ (see, *e.g.*, Frölich and Lechner, 2010) and likewise $\tau_{\text{LATU}} = \frac{\text{E}[Y(D(1))-Y(D(0))|Z=0]}{\text{E}[D(1)-D(0)|Z=0]}$; also, $\pi_z = \text{E}\left[D(1) - D(0) \mid Z = z\right]$ under Assumption UM. The remaining equalities follow from simple algebra. This completes the proof.

Proof of Corollary 3.9. Under Assumption EV, it follows from Theorem 3.7 that

$$\beta_{\text{IV}} = \frac{(1-\theta)\cdot\pi_1}{\theta\cdot\pi_0 \;+\; (1-\theta)\cdot\pi_1} \cdot \tau_{\text{LATT}} \;+\; \frac{\theta\cdot\pi_0}{\theta\cdot\pi_0 \;+\; (1-\theta)\cdot\pi_1} \cdot \tau_{\text{LATU}}. \qquad (28)$$

We also know from equation (18) that

$$\tau_{\text{LATE}} = \frac{\theta\cdot\pi_1}{\theta\cdot\pi_1 \;+\; (1-\theta)\cdot\pi_0} \cdot \tau_{\text{LATT}} \;+\; \frac{(1-\theta)\cdot\pi_0}{\theta\cdot\pi_1 \;+\; (1-\theta)\cdot\pi_0} \cdot \tau_{\text{LATU}}. \qquad (29)$$

The proof consists of three steps. First, we need to show that $\tau_{\text{LATT}} = \tau_{\text{LATU}}$ implies that $\beta_{\text{IV}} = \tau_{\text{LATE}}$. This follows immediately from equations (28) and (29) as both β_{IV} and τ_{LATE} are convex combinations of τ_{LATT} and τ_{LATU}. In fact, this implication does not even rely on Assumption EV.

Second, we need to show that $\theta = 0.5$ implies that $\beta_{\text{IV}} = \tau_{\text{LATE}}$. Indeed, if $\theta = 0.5$, then it follows from equation (28) that

$$\beta_{\text{IV}} \;=\; \frac{\pi_1}{\pi_0 + \pi_1} \cdot \tau_{\text{LATT}} \;+\; \frac{\pi_0}{\pi_0 + \pi_1} \cdot \tau_{\text{LATU}}. \tag{30}$$

Similarly, it follows from equation (29) that

$$\tau_{\text{LATE}} \;=\; \frac{\pi_1}{\pi_0 + \pi_1} \cdot \tau_{\text{LATT}} \;+\; \frac{\pi_0}{\pi_0 + \pi_1} \cdot \tau_{\text{LATU}}, \tag{31}$$

and hence $\beta_{\text{IV}} = \tau_{\text{LATE}}$.

Finally, we need to show that $\beta_{\text{IV}} = \tau_{\text{LATE}}$ implies that either $\tau_{\text{LATT}} = \tau_{\text{LATU}}$ or $\theta = 0.5$. We begin by equating the right-hand sides of equations (28) and (29). Upon rearrangement, we get

$$\frac{\theta \cdot \chi_0 \;+\; (1-\theta) \cdot \chi_1}{\theta \cdot \pi_0 \;+\; (1-\theta) \cdot \pi_1} \;=\; \frac{\theta \cdot \chi_1 \;+\; (1-\theta) \cdot \chi_0}{\theta \cdot \pi_1 \;+\; (1-\theta) \cdot \pi_0}, \tag{32}$$

where

$$\chi_z = \mathrm{E}\left[Y(D(1)) - Y(D(0)) \mid Z = z\right]. \tag{33}$$

Upon further rearrangement of equation (32), we obtain

$$\theta^2 \cdot \chi_0 \cdot \pi_1 \;+\; (1-\theta)^2 \cdot \chi_1 \cdot \pi_0 \;=\; \theta^2 \cdot \chi_1 \cdot \pi_0 \;+\; (1-\theta)^2 \cdot \chi_0 \cdot \pi_1, \tag{34}$$

which also implies that

$$(\chi_0 \cdot \pi_1 - \chi_1 \cdot \pi_0) \cdot (2\theta - 1) \;=\; 0. \tag{35}$$

For equation (35) to hold, we need either $\theta = 0.5$ or $\frac{\chi_1}{\pi_1} = \frac{\chi_0}{\pi_0}$, where the latter condition is equivalent to $\tau_{\text{LATT}} = \tau_{\text{LATU}}$. This completes the proof.

References

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

Angrist, Joshua D. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313–336.

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

Angrist, Joshua D. and Guido W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Angrist, Joshua D. and Alan B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton–Oxford: Princeton University Press.

Aronow, Peter M. and Cyrus Samii (2016): "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science*, 60, 250–267.

Bond, Simon J., Ian R. White, and A. Sarah Walker (2007): "Instrumental Variables and Interactions in the Causal Analysis of a Complex Clinical Trial," *Statistics in Medicine*, 26, 1473–1496.

Bound, John, David A. Jaeger, and Regina M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Cruz, Luiz M. and Marcelo J. Moreira (2005): "On the Validity of Econometric Techniques with Weak Instruments: Inference on Returns to Education Using Compulsory School Attendance Laws," *Journal of Human Resources*, 40, 393–410.

Donald, Stephen G., Yu-Chin Hsu, and Robert P. Lieli (2014): "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," *Journal of Business & Economic Statistics*, 32, 395–415.

Donald, Stephen G. and Whitney K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.

Evdokimov, Kirill S. and Michal Kolesár (2019): "Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects." Unpublished.

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

Frölich, Markus and Michael Lechner (2010): "Exploiting Regional Treatment Intensity for the Evaluation of Labor Market Policies," *Journal of the American Statistical Association*, 105, 1014–1029.

Hansen, Christian, Jerry Hausman, and Whitney Newey (2008): "Estimation with Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26, 398–422.

Heckman, James J. and Edward Vytlacil (2007a): "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by James J. Heckman and Edward E. Leamer, Amsterdam–Oxford: North-Holland, vol. 6B.

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

Hong, Han and Denis Nekipelov (2010): "Semiparametric Efficiency in Nonlinear LATE Models," *Quantitative Economics*, 1, 279–304.

Hull, Peter (2018): "Subtracting the Propensity Score in Linear Models." Unpublished.

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

Imbens, Guido W. and Joshua D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Ishimaru, Shoya (2019): "Beyond Linear and Homogeneous Effects: Decomposing the IV-OLS Gap in Return to Schooling Estimates." Unpublished.

Kling, Jeffrey R. (2001): "Interpreting Instrumental Variables Estimates of the Returns to Schooling," *Journal of Business & Economic Statistics*, 19, 358–364.

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

Lee, David S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Lochner, Lance and Enrico Moretti (2015): "Estimating and Testing Models with Many Treatment Levels and Limited Instruments," *Review of Economics and Statistics*, 97, 387–397.

Mogstad, Magne and Alexander Torgovitsky (2018): "Identification and Extrapolation of Causal Effects with Instrumental Variables," *Annual Review of Economics*, 10, 577–613.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters (2020): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables." Unpublished.

Rosenbaum, Paul R. and Donald B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Schochet, Peter, John Burghardt, and Sheena McConnell (2008): "Does Job Corps Work? Impact Findings from the National Job Corps Study," *American Economic Review*, 98, 1864–1886.

Semenova, Vira (2020): "Better Lee Bounds." Unpublished.

Słoczyński, Tymon (2018): "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands." IZA Discussion Paper no. 11866.

——— (2020): "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *Review of Economics and Statistics*, forthcoming.

Staiger, Douglas and James H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.