Research Article

Nick Huntington-Klein*

Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness

<https://doi.org/10.1515/jci-2020-0011>

Received May 09, 2020; accepted Aug 25, 2020

Abstract: In Instrumental Variables (IV) estimation, the effect of an instrument on an endogenous variable may vary across the sample. In this case, IV produces a local average treatment effect (LATE), and if monotonicity does not hold, then no effect of interest is identified. In this paper, I calculate the weighted average of treatment effects that is identified under general first-stage effect heterogeneity, which is generally not the average treatment effect among those affected by the instrument. I then describe a simple set of data-driven approaches to modeling variation in the effect of the instrument. These approaches identify a Super-Local Average Treatment Effect (SLATE) that weights treatment effects by the corresponding instrument effect more heavily than LATE. Even when first-stage heterogeneity is poorly modeled, these approaches considerably reduce the impact of small-sample bias compared to standard IV and unbiased weak-instrument IV methods, and can also make results more robust to violations of monotonicity. In application to a published study with a strong instrument, the preferred approach reduces error by about 19% in small ($N \approx 1,000$) subsamples, and by about 13% in larger ($N \approx 33,000$) subsamples.

Keywords: Causal Inference, Observational Studies, computational methods, economics

2020 Mathematics Subject Classification: 62D20, 91-08

1 Introduction

In order for instrumental variables (IV) estimation to identify a causal effect of interest, there are both theoretical (validity) and statistical (relevance) conditions that must hold. In applied settings, theoretical concerns about validity tend to be central. However, recent surveys of IV usage find that statistical considerations should receive more attention. Young [42] shows published IV studies often suffer from inadequate power, and Andrews, Stock, and Sun [1] find heightened sensitivity to heteroskedasticity and clustering. This occurs even though the problems of weak instruments and other forms of statistical sensitivity have been long diagnosed [33, 39] and researchers have tools for testing for weakness or addressing it.

In this paper, I keep these these statistical concerns with IV in mind while focusing on the “first stage” of estimation - the effects of instruments on their endogenous variables. Instruments may have larger or smaller effects on different individuals. I follow the consequences of heterogeneity in first-stage effects, modeling it directly to make two contributions. The first is the introduction of a simple set of linear IV estimators that improve the statistical performance of IV. The second is a demonstration of the effects identified by IV under general heterogeneity, both by typical linear two-stage least squares and by the novel estimators proposed here.

Heterogeneity in the effect of the endogenous variables in an IV setting is very well-studied [23, 27], but heterogeneity in the effect of the instruments is less so. First-stage heterogeneity is popularly understood in the framework proposed in the mid-1990s by, e.g., Angrist & Imbens [4]. Under this framework, the population

*Corresponding Author: Nick Huntington-Klein: Seattle University, United States of America;
Email: nhuntington-klein@seattleu.edu

consists of “compliers” for whom the instrument has a nonzero effect, “never-takers” and “always-takers” who are unaffected by the instrument, and “defiers” for whom the instrument has a nonzero effect of an opposite sign to the compliers. Under monotonicity (no defiers), IV estimates a local average treatment effect (LATE).¹

I present a model of effect heterogeneity in the first and second stages to show what is identified under unrestrained heterogeneity in otherwise standard settings. With one endogenous variable and one instrument, IV identifies a weighted average of all individual treatment effects, where the weights are the linear effect of the instrument on the endogenous variable. This does not match the common presentation of the IV-identified LATE as the average treatment effect (ATE) among compliers, which additionally must assume that the effect of the instrument is constant among compliers. The finding that the IV-identified LATE is generally not the average treatment effect among compliers is not novel, and in fact can be inferred from [25]. However, the simplified interpretation seems to have become common quickly, appearing by 1995 [3], and is prevalent among applied researchers.

This paper improves the statistical performance of IV by observing that the presence of observations for which the instrument has little to no effect (“never-takers” and “always-takers”) weakens the instrument and increases small-sample bias without changing the validity of the IV design or the IV estimate in expectation. Bias can be reduced by limiting the influence of these observations on estimation. There are already attempts to use this variation to strengthen the instrument using matching, as in [10], or regularization in the first stage, as in [13].

I derive the properties of two simple estimators that directly model heterogeneity in the first stage. These estimators perform standard two-stage least squares, except that the effect of the instrument is allowed to vary over groups, or is estimated at the individual level and then used as part of a sample weight. As such, these new methods should be intuitive to users of regular IV. I additionally provide software packages to aid in the usage of these methods.²

These new methods (1) identify a Super-Local Average Treatment Effect (SLATE), which is a weighted average of individual treatment effects, where weights are more strongly related to the impact of the instrument than in the LATE, (2) generally reduce noise in the IV bias term, improving statistical performance, (3) weaken the set of assumptions necessary for identification by relying on a weaker version of the monotonicity assumption in the group-interaction version of the estimator, and (4) give the researcher control over a trade-off between bias and “localness” in the weighted version of the estimator. The weighted estimator can also identify the ATE among compliers, but this relies on large samples and very accurate estimates of individual first-stage treatment effects.

The SLATE estimators are a complement to recent machine learning developments in estimating the heterogeneity of treatment effects. I estimate first-stage heterogeneity in three ways. The first two rely on no additional information or covariates. These are a naive repeated random selection (“GroupSearch”), and the Top-K τ -Path algorithm (TKTP) [12, 34–36]. Neither GroupSearch nor TKTP are capable of precisely uncovering first-stage heterogeneity, but the SLATE estimators perform well in simulation regardless. The third approach is the causal forest [8, 9, 40], which further improves performance of the SLATE estimator.

The use of modern techniques in modeling effect heterogeneity has the capacity to considerably improve estimates when combined with the SLATE estimators. I apply the new estimators in a real-world setting by replicating [2] and testing the ability to reproduce the full-sample estimate using small subsamples. In those

¹ Considerable work has been done in using instrumental variables to estimate other forms of treatment effects such as the marginal treatment effect, and in critiquing LATE for having weak economic interpretation, as in [24]. I will focus on the LATE understanding as it is common in much applied work, and relates readily to the estimand in this paper.

² The package `MagnifiedIV` can be installed in R using

```
devtools::install_github('NickCH-K/MagnifiedIV')
```

or in Stata using

```
net install MagnifiedIV,
from("https://raw.githubusercontent.com/NickCH-K/MagIVStata/master/").
```

subsamples, combining my estimators with causal forest reduces mean absolute error by about 19% in small ($N \approx 1,000$) subsamples, and by about 13% in larger ($N \approx 33,000$) subsamples.

2 Instrumental Variables with Heterogeneous Effects

2.1 One Endogenous Variable and One Excluded Instrument

In this section I demonstrate how heterogeneity in the effect of the instrument on treatment impacts the instrumental variables (IV) estimator. I use a simplified one-endogenous-variable and one-excluded-instrument setting rather than providing a general proof because the main purpose of the derivation of the weights is illustrative, and so as to drive discussion of bias. A more general derivation is not novel, and dates back at least to [25].

Consider a basic instrumental variables specification with one mean-zero endogenous variable x and one mean-zero exogenous variable z . Controls are not included. If controls are included, they have been partialled out, noting that the addition of controls will change the treatment effect weights based on the conditional residual variance in x_i and z_i after removing the effect of the controls [5, 7]. I will focus the proof on the no-controls case, with controls included only under the restrictive assumption that the treatment effect weighting introduced by controls can be ignored, aside from the group controls introduced in Section 2.2.

$$y_i = x_i\beta_i + \varepsilon_i \quad (1)$$

$$x_i = z_i\gamma_i + v_i \quad (2)$$

In other words, this is a standard instrumental-variables setup, with the corresponding general model Figure 1.

Figure 1: Basic Model Amenable to Instrumental Variables Estimation

The primary difference from a standard linear instrumental variables model is the presence of full heterogeneity in the effects of z_i on x_i (γ_i) and of x_i on y_i (β_i). Under the assumptions that $E(z'\varepsilon) = E(z'v) = E(z'\gamma) = E(z'\beta) = E(x'\gamma) = E(x'\beta) = 0$, $E(x'\varepsilon) = E(v'\varepsilon) \neq 0$, where a lack of an i subscript indicates a vector, and $E(\gamma\beta) \neq E(\gamma)E(\beta)$, then as the sample size goes to infinity, the IV estimator becomes

$$E(\hat{\beta}^{IV}) = \frac{E(\gamma'\beta)}{E(\gamma)} \quad (3)$$

as shown in Appendix A. The expected value of the IV estimator is a weighted average of the β_i s, where the weights are γ_i . The common interpretation that the LATE is the ATE among compliers only holds if γ_i is limited to two values - 0 or some constant c .

Given the weights, I turn to the small-sample bias of the IV estimator. There are two bias terms, both of which are zero in expectation but are present in finite samples:

$$\frac{\sum_i z_i v_i \beta_i}{\sum_i z_i x_i} + \frac{\sum_i z_i \varepsilon_i}{\sum_i z_i x_i} \quad (4)$$

The second of these terms is well-known from any IV derivation. The first is present because of the assumption that $E(\gamma\beta) \neq E(\gamma)E(\beta)$.

One potential means of improving the small-sample statistical properties of the IV estimator, then, is to find and remove or downweight observations with small values of γ_i , which will not heavily affect $E(\hat{\beta}^{IV})$, but should reduce noise by more in the bias term's numerator than in the denominator, increasing the share of coefficient variation driven by sampling variation rather than small-sample bias.

2.2 Modeling Variation in the Effect of the Instrument

I now consider an extension of the model in the previous section in which γ_i varies over known groups $g_i \in \{1, \dots, G\}$, and the coefficient on the instrument is allowed to vary over those groups. I examine how the identified effect and statistical performance change. Controls and group fixed effects have been partialled out in both the first and second stages. The true model is the same as in the previous section, except that the influence of these group differences are removed from v and made explicit, as in Figure 2.

Figure 2: Instrumental Variable Model with Group Heterogeneity

The estimation model becomes:

$$y_i = x_i\beta_i + \varepsilon_i \quad (5)$$

$$x_i = z_i \sum_g \gamma_g I_{gi} + v_i \quad (6)$$

where I_{gi} is an indicator function equal to 1 if $g_i = g$, and these group indicators are partialled out alongside other controls. As the sample size grows to infinity, two-stage least squares (2SLS) applied to Equations 5 and 6 identifies:

$$E(\hat{\beta}^{2SLS}) = \frac{E(\sum_i \beta_i \gamma_i \sum_g \gamma_g I_{gi})}{E(\sum_g \gamma_g^2 N_g)} \quad (7)$$

as shown in Appendix A, where $N_g = \sum_i I_{gi}$ is the number of individuals in group g . This is a weighted average of the β_i s, where the weights are $\gamma_g \gamma_i$ for the associated γ_g .

If γ_i is constant within group, this weighted average treatment effect simplifies to

$$E(\hat{\beta}^{2SLS}) = \frac{E(\sum_i \beta_i \sum_g \gamma_g^2 I_{gi})}{E(\sum_g \gamma_g^2 N_g)} \quad (8)$$

where the weights are the associated γ_g^2 for each individual, weighting the estimate more heavily on observations with high absolute γ_i values than in a LATE. I refer to any such averages as being Super-Local Average Treatment Effects (SLATE).

I then turn to statistical performance of the estimator. In a finite sample, the bias term is

$$\frac{\sum_g \hat{\gamma}_g (\sum_i z_i (v_i \beta_i + \varepsilon_i) I_{gi})}{\sum_g \hat{\gamma}_g^2 (\sum_i z_i^2 I_{gi})} \quad (9)$$

The variation in the bias term will be lower than it would be if a constant $\hat{\gamma}$ had been enforced, and the degree of reduction will be related to how different the γ_g terms are. See Appendix A; this proof does not rely on constant effects within groups.

An important feature to point out here is that in regular 2SLS, the weights are γ_i . If all γ_i s are positive, then all weights are positive. If all γ_i s are negative, then the negative term cancels out and all weights are positive. But if there is a mix of positive and negative terms across the sample (i.e. monotonicity fails), then some weights will be negative. A causal effect of interest still can be identified under weaker versions of this monotonicity assumption, such as if the negative weights can be cancelled out by positive weights on equally-sized treatment effects [18], but in general a combination of positive and negative weights is not considered to produce an estimate of interest.

However, with the $\gamma_i \gamma_g$ weights given by the SLATE method, as long as γ_i has the same sign as the associated γ_g , then any negative terms will be multiplied by another negative term, producing a positive. Instead of monotonicity needing to hold across the whole sample, using the group-based SLATE estimator, monotonicity needs only to hold within group so that all weights are positive. This finding is similar to two other papers that refine the monotonicity assumption by showing that it need only hold within subsets of the population. [17] show that monotonicity need hold only within subranges of the potential outcome space. More similar to this paper is [38], who show that after stratifying on covariates, monotonicity need only hold stochastically within those strata to identify a causal effect of interest. In these two papers as well as the present one, monotonicity is relaxed on the level of the whole sample, but must be maintained in local groups. The difference is in how those local groups are defined and empirically identified.

Allowing the effect of the instrument to vary over groups serves three purposes: it generally reduces bias, it weakens the monotonicity assumption to be monotonicity-within-group, and it increases the weight of the estimator on the β_i s associated with high γ_i values. In other words, it increases the “localness” of the estimate. This implies, in instrumental variables designs, a tradeoff between bias and localness.

2.3 Weighted IV under Full Information

Here I consider a modification of the IV estimation from the earlier section in which weights are included. Consider a diagonal matrix of weights W with $w = \{w_1, w_2, \dots\}$ on the diagonal such that $Cov(w, z) = 0$.

Following the proofs in Section 2.1, the weighted IV estimate $\hat{\beta}^{WIV}$ identifies

$$E(\hat{\beta}^{WIV}) = \frac{E((WW\gamma)' \beta)}{E(WW\gamma)} \quad (10)$$

This is a weighted average of the β_i s, where the weights are $w_i^2 \gamma_i$. Under the assumption that γ is known, and that weights are chosen such that $Cov(w, z) = 0$ and $Cov(w, \gamma) > 0$, this should reduce small-sample bias relative to two-stage least squares.

Two weighting functions likely to satisfy $Cov(w, z) = 0$ and $Cov(w, \gamma) > 0$ may be of particular interest.

The first is an indicator function $w_i = I(\gamma_i \neq 0)$. This will not change the expected value of the estimand, since observations with $\gamma_i = 0$ already receive a weight of 0 on their β_i . Many researchers already follow this weighting scheme by including data only from regions, periods, etc., where the instrument would be likely to have an effect. This can be extended such that $w_i = 0$ when γ_i indicates a defier, which restores the LATE interpretation of the estimator even if monotonicity is violated.

The second is $w_i = (F_{\gamma_i})^p$ for some $p \neq 0$, (and $w_i = 0$ if $F_{\gamma_i} = 0$ and $p < 0$) where $F_\gamma = (N - k)Var(\hat{x}|\gamma_i)/Var(x - \hat{x})$ is a first-stage F -statistic modified such that the numerator uses the variance of predicted values generated as though $\gamma = \gamma_i$ for the whole sample. In the single-instrument setting this is equivalent to $w_i = |\gamma_i|^{4p}$. This weighting scheme has the benefit of working even if γ is often small but nonzero,

and being easily applied in a multiple-instrument setting. Further, it gives the researcher some control over a bias-localness tradeoff.

With $p = 1/4$, the identified estimate in a single-instrument setting has $|\gamma_i|\gamma_i$ weights, which is conceptually similar to the $\gamma_g\gamma_i$ weights achieved by allowing the effect of the instrument to vary over groups. This makes $p = 1/4$ a natural weighting choice.

Another natural choice is $p = -1/4$ (and $w_i = 0 \forall \gamma_i = 0$). When $p = -1/4$, if the sign of γ_i is constant (no defiers), then in the single-instrument setting the treatment effect estimate uses weights $|\gamma_i|^{-1}\gamma_i = 1 \forall \gamma_i \neq 0, 0 \forall \gamma_i = 0$. In other words, $p = -1/4$ identifies the ATE among compliers, matching the standard colloquial interpretation of the LATE.

In a single-instrument setting, with $w_i = (F_{\gamma_i})^p$ weights the small-sample bias term is

$$\text{bias} = \left(\frac{\sum_i (f\gamma_i^2)^p z_i (v_i\beta_i + \varepsilon_i)}{\sum_i (f\gamma_i^2)^p z_i x_i} \right) \equiv \zeta \quad (11)$$

where $f = (N - k)\text{Var}(z)/\text{Var}(M_z x)$ and M_z is the z elimination matrix. Appendix A shows that the impact of small-sample bias is usually declining in p , with improvements more likely in small samples.

These results are dependent upon using known values of γ_i . I provide no proof here on the relationship between the precision of $\hat{\gamma}$ and the small-sample bias properties of the weighted SLATE estimator.

In sum, the weighted version of the SLATE estimator, relative to the version using a first-stage group interaction, is less certain to reduce variation in the bias term, and these proofs rely on perfect estimation of $\hat{\gamma}$. On the other hand, it offers an amount of control over the bias-localness tradeoff that the group version does not. The following simulation will provide one context in which to test whether the special conditions necessary for the weighted estimator to improve performance hold.

3 Feasible Estimators

The previous sections present estimators that rely either on knowledge of γ_i , or a set of groups over which γ_g varies. In real data, this information is generally not available.

There are many well-known methods for modeling variation in an effect using observed variables, such as with interaction terms or random slopes. There are also recent developments in machine learning for modeling heterogeneous treatment effects like causal forest. Any such approach would allow the group-based method in Section 2.2 to be performed. Alternately, any method that models γ_i directly can be used to follow the weighting method in Section 2.3, or to combine groups and weights together. Since the SLATE estimators can include controls for group identity in both stages, these approaches do not require a validity assumption for group identity.

For any such method, there is the potential concern that they will introduce bias either via overfitting or by inducing some correlation with second-stage error term v and invalidating the instrument. However, this is not a major issue.

The overfitting concern is valid, but only for the first stage: the estimate of the relationship between x and z will be overfitted unless the method chosen for modeling effect heterogeneity is shown not to overfit. But for the purposes of IV, the goal is to extract all variation in x statistically explained by z , not theoretically explained by z ; there is no particular reason that this statistical explanation needs to generalize past the present sample (see, e.g., [13]). Overfitting is acceptable.

The second concern, that modeling effect heterogeneity might invalidate the instrument, would require that z be invalid in the first place. At least in the Section 2.2 methods, the grouping structure is to be partialled out or controlled for, and so if Figure 2 requires an arrow between g and z , the instrument is still conditionally valid. For GroupSearch to invalidate the instrument, it would need to be the case that $z_i \sum_g \gamma_i I_{gi}$ is related to v while z_i is not.

It is possible that if z_i is invalid for some subgroup, and $|\gamma_g|$ is large for that subgroup, then GroupSearch could worsen the effects of invalidity by weighting that subgroup more heavily. But this requires that z_i al-

ready be invalid. As long as Figure 2 is accurate and there is no open path between z_i and v_i , this should not be possible.

The following simulation will focus on two methods for estimating first-stage heterogeneity that do not rely on covariates to model γ_i . I do this so that I will not confuse a test of the effectiveness of the estimators with success in selecting first-stage mediators. In fact, both methods only do a mediocre job at uncovering the underlying true first-stage heterogeneity, as will be discussed in Section 4.2. Despite this, the SLATE estimators still perform well. I will introduce the use of causal forest to model γ_i more accurately in Section 5.

The first method, GroupSearch, selects a number of groups and a number of iterations (100 in this paper). In each iteration, it assigns groups at random and estimates the first stage. Then, for each sample, it selects the set of groups in which the first-stage F-statistic is highest.

The second method is the Top-K τ -Path search, or TKTP [12, 34–36]. Given two variables (x and z in this case, after partialing out), TKTP is an algorithm designed to find a subset of the data in which there is a positive relationship between x and z .

TKTP uses the concordance of the ranks between the two variables. Kendall's τ is the proportion of observation pairs that are concordant ($x_i > x_j$ and $z_i > z_j$, or $x_i < x_j$ and $z_i < z_j$). A higher τ indicates a stronger positive relationship between x and z . TKTP arranges observations in order such that, if $\tau(i)$ is τ calculated using the first i observations in that order, $\tau(i)$ is decreasing. In other words, it sorts the observations by their contribution to a positive association. Given ties, the ordering may be non-unique.

Using the τ -path order, the algorithm generates the null distribution of the τ -path under no association, and identifies a stopping parameter j where the τ -path differs from the null distribution, locating a subset for which their association is statistically significantly different from the null.

In the simulation, TKTP is run twice, once on x and z to separate out a group with positive association, and once on x and $-z$ to separate out a group with negative association.

There are two minor feasibility issues with the use of TKTP. First, because there is some randomness injected in the algorithm, it is possible that the same observation may end up in both positive and negative groups, in which case it is assigned to neither. Second, under current implementations (partially drawn from [15]), it is computationally slow, and may not be usable for very large data sets, or at least not if it needs to be run thousands of times as in this simulation.

4 Simulation

I test the properties of the SLATE estimators under simulated-data settings, beginning with a setting where all IV assumptions are satisfied. Data simulation centers around the data-generating process:

$$y_i = x_i\beta_i + 2w_i + \varepsilon_i \quad (12)$$

$$x_i = z_i\gamma_i + w_i + v_i \quad (13)$$

$$z_i, w_i, \varepsilon_i, v_i \sim N(0, 1) \quad (14)$$

where w_i is an unobserved confounding factor. β_i and γ_i are constructed to be related. I encode four groups of equal size into the data: A, B, C, and D. For these groups, respectively, $\beta = \{1, 2, 3, 4\}$ and $\gamma = \{0, .075, .15, .223\}$.

These exact numbers are chosen such that the expected OLS bias is 1, and the median first-stage F-statistic is 10 at a simulated sample size of 1600. I generate 1000 simulated samples with $N = \{100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600\}$ observations each. In each sample I calculate 2SLS, as well as different versions of the SLATE estimators, by constructing groups with GroupSearch (GS) and Top-K τ -Path (TKTP) for the group-based version of the SLATE estimator. TKTP is not implemented for sample sizes above 1600

due to computational limitations. Then I use those groups to estimate γ_{gs} to use for weights with $p = 1/4$ for the weighting version of the SLATE estimator. I compare estimates to the true LATE and SLATE given the formulae in Sections 2.1 and 2.2.

4.1 Basic Simulation

Figure 3 shows performance using feasible estimation, taking as known only the number of underlying groups for use with GroupSearch (4). In results shown in the Online Supplement, performance improves further with additional groups. The SLATE estimator with GroupSearch-selected groups improves upon 2SLS by about 50% at the $N = 1600$ (first stage $F = 10$) point. Top-K τ -Path underperforms relative to GroupSearch.

In general, the SLATE group-based estimator considerably outperforms 2SLS at smaller sample sizes, and is very similar to 2SLS at large sample sizes. The weighted versions do not perform as well.

Figure 3: Performance Using Feasible Estimation

Deviation is relative to parameter identified in expectation.

Bootstrap standard errors are higher than for OLS, as shown in Figure 4. But they are in most forms smaller than 2SLS at small sample sizes, and similar at large sample sizes.

Performance is similar using $\gamma_i \sim U[0, 1/4.5]$, which is chosen to retain treatment effect averages with the original DGP.³ See the Online Supplement, which also shows the levels of performance improvement for different distributions of γ_i and joint distributions of γ_i and β_i .

The performance of the SLATE estimators are not driven by the difference between the SLATE and the LATE. Figures available in the Online Supplement measure absolute deviation relative to the true LATE identified by 2SLS, and relative to the ATE among compliers, respectively (which is not identified by any model on the graph). SLATE estimators continue to outperform 2SLS, especially in small samples.

The SLATE estimators offer improved performance compared to 2SLS under idealized conditions. However, these conditions will not hold universally. In the following sections, I create data that violate standard IV

³ In this and every other simulation using a continuous distribution for γ_i , γ_i is sorted such that the lowest quartile of γ_i values are in Group A, the next quartile is in Group B, and so on, inducing a relationship between γ_i and β_i .

Figure 4: Standard Error Using Feasible Estimation

assumptions to check whether the SLATE estimators may be especially vulnerable to these violations relative to 2SLS. I also compare SLATE to other estimators with attractive small-sample properties.

4.2 Monotonicity

In cases where monotonicity is violated, the 2SLS estimand is not of particular interest, as it contains negative weights. This can lead to an estimate that misrepresents the local average treatment effect, and may even have an opposite sign to all individual treatment effects [14]. This sensitivity to monotonicity violations also applies to the SLATE estimators unless the subsample of “defiers” can be identified for each instrument and the effect of the instrument is allowed to be different for that group. To test the sensitivity of the SLATE estimators to monotonicity violation, I repeat the DGP from Section 4 except that $\gamma_i \sim U[-1/9, 3/9]$.

Neither feasible method for identifying groups used here, GroupSearch and TKTP, accurately identifies defiers. “No relationship” was the modal group assigned by TKTP to true-negative observations across all sample sizes. Excluding “No relationship”, the modal group assigned to true-negative observations was “negative” about 40% of the time in small samples, increasing to 50% for the largest samples. In GroupSearch, the modal group assigned for true-negative observations is the lowest- $\hat{\gamma}_g$ group (out of four groups) about 25% of the time in small samples, up to 30% of the time in the largest samples. The highest- $\hat{\gamma}_g$ group was the least likely to be the modal group assigned to true-negatives, although this still did occur about 20% of the time.

Despite misclassifications, classification is a considerable improvement over no classification. In Figure 5 the SLATE estimators perform better under violations of monotonicity than 2SLS does, and the improvement is to a greater degree than in Figure 3. At the $N = 1,600$ point, for example, the GroupSearch approach reduced mean absolute deviation in Figure 3 by 30% relative to 2SLS. In Figure 5 there is instead a 57% improvement. Still, given the weakness of GroupSearch and TKTP in identifying defiers, in cases where non-monotonicity is likely, heterogeneity should be modeled using covariates likely to actually locate defiers.

Figure 5: Performance under Violation of Monotonicity

Deviation is relative to LATE or SLATE, as appropriate, with all-positive weights.

4.3 Invalidity

IV relies on a validity assumption for consistency. It is possible that the nature of the SLATE estimators, which attempt to maximize the influence of the instruments, may make the estimate more sensitive to validity violations, as described in Section 3. I present simulations that test two kinds of minor violations of validity.

First, I induce a relationship between w_i and z_i :

To test for the impact of minor violations of validity, I generate z_i as

$$z_i = .2w_i + \zeta_i; \zeta_i \sim N(0, 1) \quad (15)$$

The results of this simulation can be seen in Figure 6. Under this violation, all IV variants converge to a higher level of deviation than in Section 4.1, which is to be expected since the estimators are inconsistent. But at each sample size, the SLATE estimators continue to outperform 2SLS. Under the violation of validity, there is less deviation for the SLATE estimators at small sample sizes at large sample sizes.⁴

In addition to standard violations of validity, the SLATE estimators introduce the possibility that γ_i will be related to the second-stage error term. If this occurs, then using individualized γ_i values to predict x_i will violate validity. To test this, I return z_i to its usual $z_i \sim N(0, 1)$, and generate γ_i as

$$\gamma_i = \phi_i + .05(w_i - \min(w_i)) / \max(w_i) \quad (16)$$

where $\phi_i \sim U[0, 1/4.5]$. The results of this simulation are in Figure 7. Under this violation, the SLATE estimators are still no worse than 2SLS, and are considerably better for very small samples, but the SLATE estimators and 2SLS reach similar levels of mean absolute deviation at smaller sample sizes than in the basic simulation in the main paper, around 1600 observations rather than 6400.

⁴ Performance actually worsens in larger samples here for GroupSearch. This is because, in smaller samples, it is likely that some of the small-sample variation between groups picked up by GroupSearch is unrelated to the true underlying invalidity. As samples increase, GroupSearch picks up this invalid variation more accurately.

Figure 6: Performance With Validity Violation

Deviation is relative to parameter identified in expectation.

Figure 7: Performance with Validity Violation in γ_i

Deviation is relative to parameter identified in expectation.

4.4 Clustering

As demonstrated in [42], 2SLS is particularly sensitive to the presence of clustering and heteroskedasticity, and when i.i.d. is violated, estimates may be considerably more noisy. Following [42], I randomly assign each

observation to be one of ten clusters (allowing variation of the A, B, C, D groups within cluster). Then I modify the DGP such that

$$x_i = z_i \gamma_i + \lambda_c (\eta_c + w_i + nu_i) \sqrt{2} \quad (17)$$

$$y_i = x_i \beta_i + \lambda_c (\eta_c + 2w_i + \varepsilon_i) / \sqrt{2} \quad (18)$$

where λ_c is a randomly selected z_i value from cluster c , and η_c is a randomly selected ε_i value from cluster c . λ_c and η_c are the same for all members of cluster c .

Figure 8: Performance under Clustering

Deviation is relative to parameter identified in expectation.

Figure 8 shows the results of the simulation. The SLATE estimators still offer improved performance over 2SLS in this version, although the degree of improvement is muted, with the estimators converging to similar levels of performance at smaller sample sizes. The SLATE estimators may be harmed more in relative terms by clustering than 2SLS is. However, the SLATE estimators still outperform 2SLS in this clustered setting.

4.5 Other Weak-Instrument Methods

I compare the performance of the SLATE estimators to Jackknife Instrumental Variables Estimator (JIVE) [21] and to the [19, 26] implementation of Limited-Information Maximum Likelihood LIML, using the DGP from Section 4. I set the Fuller tuning parameter α in two ways: $\alpha = 1$ for unbiasedness, as “Fuller (1)”, and $\alpha = 4$ for minimum mean squared error, as “Fuller (4)”. Results are shown in Figure 9.

The performance of JIVE is fairly weak in the given setting, not outperforming even 2SLS. The Fuller (1) implementation of LIML, however, has similar performance to the SLATE estimators, and outperforms the Top-K τ -Path variant. Fuller (4) outperforms all SLATE estimators in mean absolute bias. However, it does return a biased result [19], and is not designed for use in cases where the instrument has heterogeneous effects, implying a tradeoff between the two estimators.

This simulation does not consider many-instrument, many-controls, or heteroskedastic contexts where LIML methods may perform more or less effectively - in particular, Fuller (4) assumes homoskedasticity, al-

Figure 9: Comparison of SLATE Estimators to Other Weak-Instrument Methods

Deviation is relative to parameter identified in expectation.

though there are heteroskedasticity-robust variants such as [22]. Additionally, this analysis ignores the interpretability issues of LIML under heterogeneous treatment. [29] shows the estimate may not be in the convex hull of local average treatment effects. There are many other small-sample robust estimators that could be tried.

4.6 Recovering the ATE Among Compliers

As discussed in Section 2.3, while 2SLS does not generally identify the ATE among compliers, the ATE among compliers can be recovered if there are no defiers, and the SLATE weighting estimator uses a weighting scheme where $w_i = 0 \forall \gamma_i = 0$ and $w_i = (F_{\gamma_i})^{-1/4}$ otherwise.

Using the original DGP from section 4, the ATE among compliers is $(2 + 3 + 4)/3 = 3$, and the IV-identified LATE is $(.075 * 2 + .15 * 3 + .223 * 4)/(.075 + .15 + .223) = 3.33$. I present deviation from the ATE among compliers.

I estimate the model three ways, as shown in Figure 10: using 2SLS, using an infeasible weighted estimator that uses the known true γ_i values, and using GroupSearch with four groups to estimate the γ_i values, setting weights to 0 for $\hat{\gamma}_i \leq 0$.

The weighted SLATE method with $p = -1/4$ comes closer to the ATE among compliers than 2SLS. However, this only works with large sample sizes, and even then only when the true γ_i values are known. This only working with large samples follows from the Section 2.3 finding that decreasing p from 0 in 2SLS to $-1/4$ in the weighted SLATE estimator should increase bias at small sample sizes. So, this method offers promise for uncovering the ATE among compliers, but only if samples are large and very accurate estimates of the γ_i s can be made.

Figure 10: Deviation from ATE Among Compliers

5 Application

In this section I demonstrate the real-world applicability of the SLATE grouping estimator by replicating Angrist, Battistin, & Vuri [2] (ABV). ABV looks at the effect of class size on student test scores, finding that much of the positive effect of smaller class sizes in Italy may be because it is easier for teachers to manipulate test scores in smaller classes. The paper identifies the effect of class sizes using a combination of the presence of randomly-assigned test monitors and class-size-maximum rules similar to the well-known Maimonides rule from [6].

ABV offers a useful setting for replication in this paper. First, ABV allows me to demonstrate the use of the SLATE estimator in a multiple-instrument setting.⁵ Second, the sample is large enough that I can demonstrate the small-sample properties of the estimator by selecting subsamples of different sizes. Third, the instrument in ABV is very strong, and so replication will demonstrate that the usefulness of the SLATE estimators is not limited to cases of weak instruments.

Fourth, the instrument should have a heterogeneous effect. While monotonicity seems likely to hold, adherence to the class-size rule is not perfect. Compliance can be graphically shown to vary with enrollment, and presumably varies by other factors as well. Figure 2b in ABV demonstrates variation in adherence to the rule.

I focus first on replicating ABV Table 6, which regresses math and Italian language scores on class size, with class sizes predicted by the class-size-maximum rule as an instrument, both of which are interacted with an indicator for being monitored. Estimation uses 2SLS with standard errors clustered at the *school* $\times$ *grade* level. A long list of controls are included, matching the controls used in the original paper. Analysis is performed separately by region.

Table 1 Panel A replicates ABV Table 6. Panel B uses GroupSearch separately for monitored and non-monitored contexts, interacting each set of resulting groups with the associated instrument. The use of five groups is arbitrary, and I also consider ten groups. I use GroupSearch and the grouping estimator only here because the sample is too large to feasibly use TKTP, and Section 4 showed that the weighting estimator does not perform well in idealized settings.

⁵ Groupings are generated for each instrument separately.

In Panel C, I use “honest” causal forests to estimate a first-stage effect for each individual, allowing the effect to vary with all covariates [9]. Because overfitting is not a concern, as discussed in Section 3, I generate individual treatment effect estimates for the full sample rather than using a holdout. I then split effects into five quintiles to perform the group estimator for each instrument.

2SLS and the GroupSearch-based SLATE estimator give very similar results in this context. This is to be expected for GroupSearch given the large sample size, and the fact that the groups are selected at random - if variation between groups is small, then the SLATE estimator in expectation approaches the LATE. The version using causal forest groupings differs from the original results by more, likely due to an improved ability to find groups with different treatment effects, but still are very similar.

The weak-instrument test F-statistics worsen for both SLATE estimators. This is because the proportion of variance explained by the instruments is only somewhat higher in the SLATE estimators than in 2SLS. With five times as many instruments, the first-stage F-statistics in the SLATE estimations are slightly more than 1/5 as large as in 2SLS.

I focus on the All Italy math score results from ABV Table 1. There are $C = 28,546$ clusters in the original data. I produce 1,000 cluster bootstrap samples for each sample size of $\{2^{-8}C, 2^{-7}C, \dots, 2^{-1}C, C\}$ clusters. I perform 2SLS, and then the SLATE grouped estimator using five-group GroupSearch, ten-group GroupSearch, and causal forest quintiles on each sample. If, following the concerns of [42], any of the estimation methods is particularly sensitive to the removal of certain clusters, this will be apparent in the results. Because of the slow speed of estimating causal forests at large samples and the need to run causal forest thousands of times (unlike in Table 1, where a large-sample causal forest is performed), I only perform causal forest estimation for sample sizes up to $2^{-3}C$.

For each cluster bootstrap sample I calculate mean absolute deviation from the full-sample parameters generated with the same method. Figure 11 shows the results for math scores in monitored classrooms. The figure for non-monitored classrooms looks similar and is in the Online Supplement. Both figures show convergence for both endogenous variables towards the parameters they identify. If the target is instead the 2SLS result in Table 1, the relative performance of the estimators does not change.

Figure 11: Performance in Replication of ABV Monitored $\times$ Class Size

Deviation is relative to the full-sample estimate in Table 1. Causal Forest is only estimated for smaller samples due to computational limitations.

Table 1: Replication of ABV Table 6 Without and With the SLATE Estimator

	Original Results					
	Math Scores			Language Scores		
	All Italy	N/Center	South	All Italy	N/Center	South
Class Size ×	−0.035	−0.039*	−0.035	−0.031	−0.021	−0.048
Monitored	(0.024)	(0.021)	(0.060)	(0.019)	(0.017)	(0.048)
Class Size ×	−0.066***	−0.042**	−0.143***	−0.042**	−0.021	−0.098**
Not Monitored	(0.021)	(0.018)	(0.053)	(0.016)	(0.014)	(0.042)
Monitored	−0.174***	−0.082**	−0.395***	−0.103***	−0.055*	−0.228***
	(0.041)	(0.038)	(0.096)	(0.033)	(0.030)	(0.076)
Weak IV F Mon.	44691	34569	12093	44691	34569	12093
Not Monitored	23072	19291	5552	23072	19291	5552
SLATE Grouping Estimator with GroupSearch (5 Groups)						
	Math Scores			Language Scores		
	All Italy	N/Center	South	All Italy	N/Center	South
Class Size ×	−0.036	−0.038*	−0.038	−0.030	−0.021	−0.047
Monitored	(0.024)	(0.021)	(0.060)	(0.019)	(0.017)	(0.048)
Class Size ×	−0.066***	−0.041**	−0.144***	−0.042**	−0.021	−0.097**
Not Monitored	(0.021)	(0.018)	(0.053)	(0.016)	(0.014)	(0.042)
Monitored	−0.173***	−0.081**	−0.389***	−0.104***	−0.055*	−0.229***
	(0.041)	(0.038)	(0.096)	(0.033)	(0.030)	(0.076)
Weak IV F Mon.	8943	6919	2423	8948	6917	2419
Not Monitored	4615	3858	1111	4615	3858	1110
SLATE Grouping Estimator with Causal Forest Quintiles (5 Groups)						
	Math Scores			Language Scores		
	All Italy	N/Center	South	All Italy	N/Center	South
Class Size ×	−0.029	−0.041**	−0.034	−0.024	−0.022	−0.040
Monitored	(0.023)	(0.021)	(0.060)	(0.019)	(0.017)	(0.048)
Class Size ×	−0.069***	−0.042**	−0.142***	−0.042**	−0.024	−0.087**
Not Monitored	(0.021)	(0.018)	(0.053)	(0.016)	(0.014)	(0.042)
Monitored	−0.191***	−0.076**	−0.393***	−0.117***	−0.056*	−0.223***
	(0.039)	(0.037)	(0.094)	(0.031)	(0.030)	(0.075)
Weak IV F Mon.	9435	7081	2465	9417	7024	2468
Not Monitored	4793	3933	1140	4795	3903	1139

Note: Panel A replicates ABV [2] Table 6. Panels B and C repeat that analysis using the grouped SLATE estimator, with GroupSearch and causal forest to identify groups, respectively. See Section 5. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

In Figure 11, all SLATE estimators tested outperform 2SLS in smaller samples. The version using causal forest to generate first-stage groups continues to outperform 2SLS in larger samples. This makes sense given that the performance gains are tied to the ability to find groups over which γ_g varies, and causal forest should be more successful at that task than GroupSearch.

This replication shows the power of the SLATE estimators to improve performance even in this setting where samples are relatively large and typical diagnostics would not warn about weak instruments. In Figure 11, at 1/2 of the original clusters ($C = 14, 273, N \approx 110,000$), there is a small difference: the GroupSearch methods outperform 2SLS by about .4%. At 1/8 of the original clusters ($C = 7, 136, N \approx 33,000$), the GroupSearch methods improve upon 2SLS by 1.3%, and the causal forest approach improves upon 2SLS by 13.4%. At the smallest sample tested ($C = 111, N \approx 1,000$), mean absolute deviation in the GroupSearch

estimator is 22.6% lower than the mean absolute deviation in 2SLS for the 10-group GroupSearch method, and 19.2% lower for causal forest.

6 Conclusion

Instrumental variables (IV) is at an odd point in its history. It seems that economists in general have grown more skeptical about instrument validity assumptions, or at least have shifted to higher standards for instruments. For example, compare [31] to [37] on the use of rainfall as an instrument. In addition to the theoretical assumptions necessary to use IV, the statistical properties of IV are also a point of concern. Recent meta-analytic studies on IV as it is performed, [42] and [1], show that studies often suffer from inadequate power, and heightened sensitivity to heteroskedasticity and clustering.

The reconstruction of IV necessarily must proceed on both fronts. Theoretical improvements, and thus improvements in identification, can come from stricter evaluation of exclusion restrictions, either theoretically or from the use of joint validity-monotonicity tests in contexts where those tests can be applied [11, 32]. There is also potential in the series of new IV estimators that weaken the reliance on validity assumptions [28, 30, 41]. Versions of the IV estimator that make statistical improvements under small samples or weak instruments already exist, especially under homoskedasticity, but are not applied anywhere near a universal scale, even in top publications (see [1] for a review, as well as [16] for the related literature on estimation with many weak instruments). Statistical improvements can come from more consistent application of methods robust to weak instruments.

I introduce an estimation approach that incorporates heterogeneity in the first-stage estimate. This approach changes what is identified by IV, identifying a super-local average treatment effect (SLATE) that weights observations with strong first-stage effects more strongly than they are already weighted in a local average treatment effect (LATE). These SLATE estimators may be worth the change in identification because they have superior statistical performance relative to traditional two-stage least squares, achieved by downweighting the impact of observations where the instrument has little impact, which add noise to the small-sample bias term.

While the ATE is generally considered the preferred estimate, it is not clear that the SLATE estimated in this paper is of less policy relevance than the LATE, which already sacrifices generalizability in favor of reduced bias, and so a more precisely-estimated SLATE may be preferable to a more-biased LATE. In policy settings where an ATE is desired, neither estimator properly identifies the ATE, but the SLATE reduces bias and provides information on which observations are being most heavily weighted, allowing for a more careful consideration on how the (S)LATE and ATE might differ. In settings where the effect of interest is intervention-based, such as in a marginal treatment effect or in preparation for a future policy, then the SLATE weights more heavily the effects of individuals who are responsive to the instrument. If the instrument is itself an assignment mechanism, or if responsiveness to the instrument varies because of latent differences in how malleable the endogenous variable is for different individuals, then heavier weights on more malleable individuals will get closer to the desired effect, even though a true marginal treatment effect is not identified. Aside from all of this, there may be many settings in which, despite the presence of heterogeneous effects, the SLATE and LATE are simply not that different, and so bias reduction does not come at a cost of changing the estimand greatly. Section 5 is an example of this.

However, if researchers do prefer the LATE to the SLATE, they should be aware that including an interaction term between the instrument and a group identifier, which is a relatively common practice, produces a SLATE rather than a LATE. This issue of accidentally producing a SLATE applies to the recent popularity of methods that use many interactions in the first stage, with regularization to select between them, implemented in [13] and [20], although the particular SLATE identified in these cases will be more difficult to derive if the group interactions overlap.

The group-interaction variant of the SLATE estimator, which outperforms the weighting variant, has the benefit of being extremely simple. It can be implemented in any linear IV context without modifying the es-

timization method or code except to add a method for identifying groups. As opposed to other small-sample robust IV methods, researchers may be more willing to implement a SLATE estimator for this reason. The group variant of SLATE is simple enough that other papers have already implemented it using group covariates already in their data, although to my knowledge no researcher doing so has reported estimating a SLATE, of which they should be aware.

The simulations in this paper find considerable success for the group SLATE estimator even under assumption violations. Researchers can achieve improved performance with a SLATE estimator even if the group-identification method performs no better than GroupSearch, which operates via naive random repeated classification, although results will improve further using causal forest or another method that uses covariates.

SLATE estimators are also capable of improving robustness to monotonicity violations. Standard IV estimation, and its small-sample-robust variations, are not robust to violations of monotonicity, and rely on assuming that monotonicity holds. The group-based SLATE estimators rely on a weaker version of the monotonicity assumption, where monotonicity must hold only within groups, similar to other forms of localized monotonicity assumptions as in [17, 38].

In addition to these general benefits of using a group-interaction SLATE estimator, right now is an opportune time to emphasize the modeling of first-stage heterogeneity. The SLATE estimator is most powerful when heterogeneity in the IV first stage is well-understood. While hierarchical modeling has long allowed for effect heterogeneity to be closely modeled, this approach relies on random-effects assumptions that economists have been skeptical of, and it is not common to use hierarchical modeling in the first stage of an IV model. Recent developments overlapping with computer science have improved the ability to estimate heterogeneity in treatment effects. Causal forest considerably improves performance of the SLATE estimator in an applied context, and other work in machine learning on treatment effect heterogeneity is underway.

Of course, this paper's method only improves IV estimation along the lines of relevance and monotonicity. It does not address validity, and while its improved small-sample properties cancel out some of IV's weakness to clustering in simulation, the estimator does not directly address the issue. Improving small-sample properties does not matter much if validity assumptions are looked upon with increasing skepticism. Still, IV is also used in cases where validity may be considered more defensible, like in fuzzy regression discontinuity or imperfect random assignment. Here an improvement in statistical performance can be combined with solid theoretical assumptions. Future work combining first-stage heterogeneity with the novel crop of IV methods more robust to violations of validity would be valuable.

Notes

⁶This does not require that the variance of z_i within group be unrelated across groups to γ_g since all between-group calculations in the denominator are additively separable.

References

- [1] Andrews, I., J. H. Stock, and L. Sun (2019): "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11, 727–753.
- [2] Angrist, J. D., E. Battistin, and D. Vuri (2017): "In a small moment: Class size and moral hazard in the Italian mezzogiorno," *American Economic Journal: Applied Economics*, 9, 216–249.
- [3] Angrist, J. D. and G. W. Imbens (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American Statistical Association*, 90, 431–442.
- [4] Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91, 444–455.
- [5] Angrist, J. D. and A. B. Krueger (1999): "Empirical strategies in labor economics," in *Handbook of Labor Economics*, volume 3, Elsevier, 1277–1366.

- [6] Angrist, J. D. and V. Lavy (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly Journal of Economics*, 114, 533–575.
- [7] Aronow, P. M. and C. Samii (2016): "Does regression produce representative estimates of causal effects?" *American Journal of Political Science*, 60, 250–267.
- [8] Athey, S. and G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.
- [9] Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized random forests," *The Annals of Statistics*, 47, 1148–1178.
- [10] Baiocchi, M., D. S. Small, S. Lorch, and P. R. Rosenbaum (2010): "Building a stronger instrument in an observational study of perinatal care for premature infants," *Journal of the American Statistical Association*, 105, 1285–1296.
- [11] Balke, A. and J. Pearl (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1171–1176.
- [12] Bamattre, S., R. Hu, and J. S. Verducci (2017): "Nonparametric testing for heterogeneous correlation," in S. E. Ahmed, ed., *Big and Complex Data Analysis: Methodologies and Applications*, Contributions to Statistics, Cham: Springer International Publishing, 229–246.
- [13] Belloni, A., V. Chernozhukov, and C. Hansen (2014): "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, 28, 29–50.
- [14] Burgess, S. and D. S. Small (2016): "Predicting the direction of causal effect based on an instrumental variable analysis: A cautionary tale," *Journal of Causal Inference*, 4, 49–59.
- [15] Caloiaro, A. (2019): "Topk tau-path," <https://github.com/acaloiaro/topk-taupath>, accessed: 2019-09-02.
- [16] Chao, J. C. and N. R. Swanson (2005): "Consistent estimation with a large number of weak instruments," *Econometrica*, 73, 1673–1692.
- [17] Dahl, C. M., M. Huber, and G. Mellace (2017): "It's never too late: A new look at local average treatment effects with or without defiers," *Discussion Papers on Business and Economics*, University of Southern Denmark, 2.
- [18] De Chaisemartin, C. (2017): "Tolerating defiance? local average treatment effects without monotonicity," *Quantitative Economics*, 8, 367–396.
- [19] Fuller, W. A. (1977): "Some properties of a modification of the limited information estimator," *Econometrica*, 45, 939–953.
- [20] Gannaway, G. (2019): "Comparative advantage in health care delivery: A machine learning approach," Unpublished Working Paper.
- [21] Ginestet, C. E. (2016): "SteinIV: Semi-parametric stein-like estimator with instrumental variables," <https://CRAN.R-project.org/package=SteinIV>, accessed: 2019-09-27.
- [22] Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012): "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics*, 3, 211–255.
- [23] Heckman, J. J., S. Urzua, and E. Vytlačil (2006): "Understanding instrumental variables in models with essential heterogeneity," *The Review of Economics and Statistics*, 88, 389–432.
- [24] Heckman, J. J. and E. J. Vytlačil (2007): "Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation," *Handbook of Econometrics*, 6, 4779–4874.
- [25] Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.
- [26] Jiang, Y., H. Kang, D. Small, and Q. Zhao (2017): "ivmodel: Statistical inference and sensitivity analysis for instrumental variables model," <https://CRAN.R-project.org/package=ivmodel>, accessed: 2019-09-27.
- [27] Kasy, M. (2014): "Instrumental variables with unrestricted heterogeneity and continuous treatment," *The Review of Economic Studies*, 81, 1614–1636.
- [28] Kippersluis, H. v. and C. A. Rietveld (2018): "Beyond plausibly exogenous," *The Econometrics Journal*, 21, 316–331.
- [29] Kolesár, M. (2013): "Estimation in an instrumental variables model with treatment effect heterogeneity," Unpublished Working Paper.
- [30] Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2015): "Identification and inference with many invalid instruments," *Journal of Business & Economic Statistics*, 33, 474–484.
- [31] Miguel, E. and S. Satyanath (2011): "Re-examining economic shocks and civil conflict," *American Economic Journal: Applied Economics*, 3, 228–232.
- [32] Mourifié, I. and Y. Wan (2017): "Testing local average treatment effect assumptions," *Review of Economics and Statistics*, 99, 305–313.
- [33] Nelson, C. R. and R. Startz (1990): "The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one," *The Journal of Business*, 63, S125–S140.
- [34] Sampath, S., A. Caloiaro, W. Johnson, and J. S. Verducci (2015): "The top-k tau-path screen for monotone association," Technical report, arXiv:1509.00549 [stat].
- [35] Sampath, S., A. Caloiaro, W. Johnson, and J. S. Verducci (2016): "The top-k tau-path screen for monotone association in subpopulations," *Wiley Interdisciplinary Reviews: Computational Statistics*, 8, 206–218.
- [36] Sampath, S. and J. S. Verducci (2013): "Detecting the end of agreement between two long ranked lists," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6, 458–471.
- [37] Sarsons, H. (2015): "Rainfall and conflict: A cautionary tale," *Journal of Development Economics*, 115, 62–72.

- [38] Small, D. S., Z. Tan, R. R. Ramsahai, S. A. Lorch, M. A. Brookhart, et al. (2017): “Instrumental variable estimation with a stochastic monotonicity assumption,” *Statistical Science*, 32, 561–579.
- [39] Staiger, D. and J. H. Stock (1997): “Instrumental variables regression with weak instruments,” *Econometrica; Evanston*, 65, 557–586.
- [40] Wager, S. (2018): “Estimation and inference of heterogeneous treatment effects using random forests,” *Journal of the American Statistical Association*, 113, 1228–1242.
- [41] Windmeijer, F., H. Farbmacher, N. Davies, and G. D. Smith (2018): “On the use of the lasso for instrumental variables estimation with some invalid instruments,” *Journal of the American Statistical Association*, 114, 1339–1350.
- [42] Young, A. (2018): “Consistency without inference: Instrumental variables in practical application,” Unpublished.

A Appendix: Proofs

Proof of Equation 3:

A standard IV estimator is calculated as:

$$\hat{\beta}^{IV} = \frac{\widehat{Cov}(z, y)}{\widehat{Cov}(z, x)} \quad (A19)$$

where N is the sample size.

$$\widehat{Cov}(z, y) = \frac{1}{N} \sum_i z_i y_i = \frac{1}{N} \sum_i z_i (x_i \beta_i + \varepsilon_i) = \frac{1}{N} \sum_i (z_i x_i \beta_i + z_i \varepsilon_i) \quad (A20)$$

$$= \frac{1}{N} \sum_i (z_i (z_i \gamma_i + v_i) \beta_i + z_i \varepsilon_i) = \frac{1}{N} \sum_i (z_i^2 \gamma_i \beta_i + z_i v_i \beta_i + z_i \varepsilon_i)$$

$$\widehat{Cov}(z, x) = \frac{1}{N} \sum_i z_i x_i = \frac{1}{N} \sum_i (z_i^2 \gamma_i + z_i v_i) \quad (A21)$$

The ratio of these two gives the small-sample estimate of $\hat{\beta}^{IV}$.

Note that all individual variables in the denominator are unrelated by assumption, including z_i^2 and γ_i . Under the assumption that z_i and γ_i have finite variance and expected value, respectively, the denominator converges in probability to a constant $E(z^2)E(\gamma)$, allowing Slutsky's theorem to be applied so that

$$E(\hat{\beta}^{IV}) = \frac{E(\frac{1}{N} \sum_i (z_i^2 \gamma_i \beta_i + z_i v_i \beta_i + z_i \varepsilon_i))}{E(z^2)E(\gamma)} \quad (A22)$$

In expectation, since $E(z' \varepsilon) = E(z' v) = Cov(z, \gamma) = Cov(z, \beta) = 0$ and, as above, z_i^2 and γ_i are unrelated, the $z_i v_i \beta_i$ and $z_i \varepsilon_i$ terms in the numerator converge in probability to 0 and this becomes

$$E(\hat{\beta}^{IV}) = \frac{E(z^2 \gamma' \beta)}{E(z^2)E(\gamma)} = \frac{E(\gamma' \beta)}{E(\gamma)} \quad (A23)$$

Proof of Equation 7:

Estimating this model by 2SLS without controls, the fitted values in the first stage are equivalent to what would arise by estimating the first stage G separate times, once for each group.

$$\hat{x}_i = z_i \sum_g \hat{\gamma}_g I_{gi} = z_i \sum_g \frac{Cov(x_i, z_i | I_{gi})}{Var(z_i | I_{gi})} I_{gi} \quad (A24)$$

where $\hat{\gamma}_g$ is the first-stage coefficient estimated for group g , and γ_g is the true mean γ_i among those in group g . The 2SLS estimator is

$$\hat{\beta}^{2SLS} = \frac{\widehat{Cov}(\hat{x}, y)}{\widehat{Var}(\hat{x})} \quad (A25)$$

The numerator and denominator can be expanded as

$$\begin{aligned} \widehat{Cov}(\hat{x}, y) &= \frac{1}{N} \sum_i \hat{x}_i y_i = \frac{1}{N} \sum_i z_i y_i \sum_g \hat{\gamma}_g I_{gi} \\ &= \frac{1}{N} \sum_i (z_i^2 \gamma_i \beta_i + z_i v_i \beta_i + z_i \varepsilon_i) \sum_g \hat{\gamma}_g I_{gi} \\ &= \frac{1}{N} \sum_g \hat{\gamma}_g \left(\sum_i (z_i^2 \gamma_i \beta_i + z_i v_i \beta_i + z_i \varepsilon_i) I_{gi} \right) \end{aligned} \quad (A26)$$

$$\begin{aligned}
 (\widehat{Var}(\hat{x})) &= \frac{1}{N} \sum_i \hat{x}_i^2 = \frac{1}{N} \sum_i \left(z_i \sum_g \hat{\gamma}_g I_{gi} \right)^2 = \frac{1}{N} \sum_g \hat{\gamma}_g^2 \left(\sum_i z_i^2 I_{gi} \right) \\
 &= \frac{1}{N} \sum_g \left(\frac{Cov(x_i, z_i | I_{gi})}{Var(z_i | I_{gi})} \right)^2 \left(\sum_i z_i^2 I_{gi} \right) = \frac{1}{N} \sum_g \left(\frac{(Cov(x_i, z_i | I_{gi}))^2}{Var(z_i | I_{gi})} \right)
 \end{aligned} \tag{A27}$$

Under the assumption that z_i has a finite variance within each group, and x_i and z_i have a finite covariance within each group,⁶ each term converges in probability to a constant, and the sum converges in probability to a constant $\sum_g \gamma_g^2 E(z_i^2 | I_{gi})$, allowing Slutsky's theorem to be applied so that

$$E(\hat{\beta}^{IV}) = \frac{E(\frac{1}{N} \sum_g \hat{\gamma}_g (\sum_i (z_i^2 \gamma_i \beta_i + z_i v_i \beta_i + z_i \varepsilon_i) I_{gi}))}{\frac{1}{N} \sum_g \gamma_g^2 E(z_i^2 | I_{gi})} \tag{A28}$$

In expectation, $E(z_i v_i) = E(z_i \varepsilon_i) = 0$ and so the $z_i v_i$ and $z_i \varepsilon_i$ terms drop out of the numerator. As a result, 2SLS identifies

$$E(\hat{\beta}^{2SLS}) = \frac{E(\sum_i \beta_i \gamma_i \sum_g \gamma_g I_{gi})}{E(\sum_g \gamma_g^2 N_g)} \tag{A29}$$

where $N_g = \sum_i I_{gi}$ is the number of individuals in group g . Notice that since $\gamma_g = \frac{1}{N} \sum_i \gamma_i I_{gi}$, the denominator is equivalent to $E(\sum_g \gamma_g \sum_i \gamma_i I_{gi})$, reinforcing the interpretation of this term as representing a weighted average of the β_i s, where the weights are $\gamma_g \gamma_i$.

Proof of Reduced Bias in Equation 9:

Compared to the bias term in Section 2.1, each term in the summation is multiplied by an additional $\hat{\gamma}_g$ in the numerator and the denominator. I rewrite the bias by pulling out what each term in the summation would be if $\hat{\gamma}$ were not allowed to vary over group:

$$\frac{\sum_g (\hat{\gamma}_g - \hat{\gamma}) (\sum_i z_i (v_i \beta_i + \varepsilon_i) I_{gi}) + \hat{\gamma} \sum_i z_i (v_i \beta_i + \varepsilon_i)}{\sum_g (\hat{\gamma}_g^2 - \hat{\gamma}^2) (\sum_i z_i^2 I_{gi}) + \hat{\gamma}^2 \sum_i z_i^2} \tag{A30}$$

Consider the variance of this bias under i.i.d.:

$$\frac{\sum_g E(\hat{\gamma}_g^2 (\sum_i z_i^2 (v_i \beta_i + \varepsilon_i)^2 I_{gi}))}{\sum_g E(\hat{\gamma}_g^4 (\sum_i z_i^4 I_{gi}))} \tag{A31}$$

By the BLUE properties of OLS, estimating the first stage separately by group will necessarily increase $\widehat{Var}(\hat{x})$. Recall that between-group differences have already been partialled out from both x and z , so Simpson's paradox does not apply here. So, taking $\hat{\gamma}_g = \gamma_g$ and assuming that the $(v_i \beta_i + \varepsilon_i)^2$ term is separable, the variation in the bias term will be lower than it would be if a constant $\hat{\gamma}$ had been enforced, and the degree of reduction will be related to how different the γ_g terms are.

In a given finite sample, these final two assumptions may not hold. Further, the reduction in bias is less likely the more noise there is in $\hat{\gamma}_g$ (i.e. the smaller the groups are). There is also always the possibility that in a given finite sample, $\hat{\gamma}_g$ may be related to $z_i^2 (v_i \beta_i + \varepsilon_i)^2$, increasing bias relative to regular IV.

⁶ This does not require that the variance of z_i within group be unrelated across groups to γ_g since all between-group calculations in the denominator are additively separable.

Proof of Impact of p on Equation 11:

Using the bias term from Section 2.1, the bias term from the weighted estimator is

$$\frac{\sum_i w_i^2 z_i v_i \beta_i}{\sum_i w_i^2 z_i x_i} + \frac{\sum_i w_i^2 z_i \varepsilon_i}{\sum_i w_i^2 z_i x_i} \quad (\text{A32})$$

With $w_i = (F_{\gamma_i})^p$ weights, where $F_{\gamma} = (N - k) \text{Var}(\hat{x}|\gamma_i) / \text{Var}(x - \hat{x})$, the derivative of the bias term with respect to p is

$$\frac{\partial \zeta}{\partial p} = \frac{\sum_{i|\gamma_i \neq 0} \log(f_{\gamma_i}^2) (f_{\gamma_i}^2)^p z_i (v_i \beta_i + \varepsilon_i)}{\sum_i (f_{\gamma_i}^2)^p z_i x_i} - \zeta \frac{\sum_{i|\gamma_i \neq 0} \log(f_{\gamma_i}^2) (f_{\gamma_i}^2)^p z_i x_i}{\sum_i (f_{\gamma_i}^2)^p z_i x_i} \quad (\text{A33})$$

Assume that $|f_{\gamma_i}^2|$ is either equal to 0 or above 1 for all i (or use a related weighting scheme where $w_i = 0 \forall |f_{\gamma_i}^2| < 1$ but otherwise $w_i = |f_{\gamma_i}^2|$).

In a finite sample, increasing p is not guaranteed to reduce bias, but a reduction will be more likely the larger the bias is. As long as the γ_i values are generally of the same sign, $\frac{\sum_{i|\gamma_i \neq 0} \log(f_{\gamma_i}^2) (f_{\gamma_i}^2)^p z_i x_i}{\sum_i (f_{\gamma_i}^2)^p z_i x_i}$ will on average be positive and above 1. If Equation A33 is dominated by the second term, then it will have the opposite sign of ζ and so increases in p will shrink the bias towards 0.

Variation in sign of the γ_i values can reduce the term below 1 and can even make it negative. For example, consider if the largest $z_i x_i$ terms in absolute value are of the opposite sign (WLOG, negative) of most of the $z_i x_i$ terms (positive). The large number of positive $z_i x_i$ terms makes $E((f_{\gamma_i}^2)^p z_i x_i)$ positive, but the additional weight given the large negative terms by $\log(f_{\gamma_i}^2)$ may make $E(\log(f_{\gamma_i}^2) (f_{\gamma_i}^2)^p z_i x_i)$ negative.

Does the second term dominate Equation A33? The second term takes the bias, reverses its sign, and, on average, scales it up, which means that it will be greater in absolute value than ζ alone. The first term takes the bias and multiplies each summative element of the bias by $\log(f_{\gamma_i}^2)$. Because γ_i is unrelated to z_i , v_i , and ε_i , it is ambiguous whether this will be greater or lesser than ζ . So while it is not guaranteed, in general, the second term should dominate and p will reduce bias. However, as ζ shrinks, variation in the second term drops relative to variation in the first term, so the first term should dominate more often than it does at small sample sizes. Since ζ decreases with sample size, the chance that increases in p worsen bias increases for larger sample sizes. As the sample size grows, the chance that a reduction in p might reduce bias increases. The proof follows very similarly if instead the variance of the bias, ζ^2 , is used.

B Simulation Subsections

Each of the below simulations use the following data-generating process as a basis, and make modifications to it.

$$y_i = x_i \beta_i + 2w_i + \varepsilon_i \quad (\text{A34})$$

$$x_i = z_i \gamma_i + w_i + v_i \quad (\text{A35})$$

$$z_i, w_i, \varepsilon_i, v_i \sim N(0, 1) \quad (\text{A36})$$

where w_i is an unobserved confounding factor. β_i and γ_i are constructed to be related. I encode four groups of equal size into the data: A, B, C, and D. For these groups, respectively, $\beta = \{1, 2, 3, 4\}$ and $\gamma = \{0, .075, .15, .223\}$. In every version of the simulation using a continuous distribution for γ_i , γ_i is sorted such that the lowest quartile of γ_i values are in Group A, the next quartile is in Group B, and so on, inducing a relationship between γ_i and β_i .

These exact numbers are chosen such that the expected OLS bias is 1, and the median first-stage F-statistic is 10 at a simulated sample size of 1, 600. I generate 1, 000 simulated samples with $N = \{100, 200, 400,$

800, 1600, 3200, 6400, 12800, 25600} observations each. In each sample I calculate 2SLS, as well as different versions of the SLATE estimators, by constructing groups with GroupSearch (GS) and Top-K τ -Path (TKTP) for the group-based version of the SLATE estimator. TKTP is not implemented for sample sizes above 1,600 due to computational limitations.

C Basic Simulation

Figure A12 shows the relative performance of the SLATE and 2SLS estimators with a uniform distribution of γ_i . Performance with a linear estimator is similar to performance using the data generating process from the basic simulation section.

Figure A12: Performance Using Feasible Estimation - Linear Specification

Figure A13 gives the extent to which the SLATE estimator outperforms the 2SLS estimator under different data generating processes. Specifically, with $\gamma_i \sim U[0, x]$ for different values of x , and $\beta_i \sim N(y\gamma_i, 1)$ for different values of y . The data generating process is otherwise the same as in the basic simulation section. Attempted values are $x \in \{2/50, 4/50, \dots, 1\}$ and $y \in \{-24/50, -22/50, \dots, 22/50, 24/50\}$. In each case, samples of $N = 1,600$ are drawn 500 times. The median bias for each estimator across all 500 samples is calculated and compared.

For all chosen values, SLATE outperforms 2SLS at the median. The scale of outperformance happens to be between 0 and 1, but this is not guaranteed. Gains do not appear to be affected by the degree of relationship between γ_i and β_i , but they are heavily affected by the distribution of γ_i . SLATE outperforms 2SLS most heavily under weak instrument conditions, which is not surprising given that this is where 2SLS bias is largest.

D Number of Groups

This paper offers relatively little guidance in selecting the number of groups for GroupSearch, or any other method that splits the sample into groups, such as how I use causal forest in the application in the main paper. Cross validation, a standard tool for selecting parameters, does not make much sense for GroupSearch where

Figure A13: Performance Using Feasible Estimation - Differing Effects Distributions

All sample sizes are 1,600, and for each combination of values, 500 samples are drawn and GroupSearch is used to feasibly estimate the SLATE estimator. Median bias among the 500 2SLS estimates minus median bias among the 500 SLATE estimates is graphed.

the groups are selected randomly. The only restriction the model does outline is that there should not be so many groups such that $\hat{\gamma}_g$ is very noisily estimated. Here I examine the extent to which this is likely to be an issue by performing GroupSearch with different numbers of groups $\{2, 4, \dots, 16\}$. The distribution of γ_i is changed such that $\gamma_i \sim U[0, 1/4.5]$ so there is not a true underlying number of groups.

Figure A14: GroupSearch Using Different Numbers of Groups

Deviation is relative to parameter identified in expectation.

At least in these simplified settings, increasing the number of groups monotonically improves performance, even at very low sample sizes where there are fewer than ten observations in each group. The tradeoff inherent in increasing the number of groups between increasing noise in $\hat{\gamma}_g$ and increasing $\text{Var}(\hat{\gamma}_g)$ has not yet reached a point where small-sample bias increases. It seems likely that the model is highly overfit with 16 groups in 100 observations, but this does not harm performance of the estimator. Performance of the SLATE estimator in previous sections and in the main paper could be improved further with the use of more groups.

E Changing Targets

Figures A15 and A16 show the relative performance of the SLATE and 2SLS estimators given shared targets. Instead of measuring bias relative to the identified parameter (SLATE and LATE, respectively), they measure bias relative to the LATE, and to the average treatment effect among compliers.

Performance of the SLATE estimator in Figure A15 in terms of absolute deviation relative to the LATE identified by 2SLS looks very similar to the main results presented in Figure 1 in the original paper. This may be surprising given that the SLATE estimator does not identify the LATE at all. Simply because the estimator is less noisy and the SLATE and LATE are near each other, its estimates are close to LATE more consistently than 2SLS estimates are in small samples.

Figure A15: Performance Using Feasible Estimation - Deviation Relative to LATE

Absolute deviation is measured relative to the true local average treatment effect identified by 2SLS.

The SLATE estimators in Figure A16 once again outperform 2SLS in small samples, simply because they are less noisy and so estimates are somewhat close to the ATE among compliers. However, both the SLATE estimators and 2SLS perform much worse in absolute terms, which is not surprising as neither identifies the ATE among compliers.

Figure A16: Performance Using Feasible Estimation - Deviation Relative to ATE Among Compliers

Absolute deviation is measured relative to the true average treatment effect among the compliers (those for whom the instrument has a nonzero effect).

F Application

See Figure A17. Also, all data for this application are available at <https://www.aeaweb.org/articles?id=10.1257/app.20160267>.

Figure A17: Performance in Replication of ABV Not Monitored × Class Size

Deviation is relative to the full-sample estimate in Table 1 in the original paper.