Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness

Nick Huntington-Klein[a,*]

[a] *California State University, Fullerton*

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 22% in small ($N \approx 1,000$) subsamples, and by about 13% in larger ($N \approx 33,000$) subsamples.

Keywords: Econometrics, Instrumental Variables, Machine Learning, Heterogeneous Effects

JEL: C26, C63, C13

*Corresponding Author. Email: nhuntington-klein@fullerton.edu. Address: 800 N. State College Blvd., Fullerton, CA, 92831.

I. 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. Published IV studies often suffer from inadequate power (Young, 2018) and heightened sensitivity to heteroskedasticity and clustering (Andrews et al., 2019). This occurs even though the problem of weak instruments and other forms of statistical sensitivity has been long diagnosed (Nelson and Startz, 1990; Staiger and Stock, 1997) and researchers have tools for testing for weakness or addressing it.

This paper provides a set of simple IV estimators that improve the statistical performance of IV by 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 model this heterogeneity directly and examine how it relates to the identification of causal effects, and to the statistical performance of IV.

Heterogeneity in the effect of the endogenous variables in an IV setting is very well-studied (e.g. Kasy, 2014; Heckman et al., 2006) but heterogeneity in the effect of the instruments less so. First-stage heterogeneity is commonly understood in the framework proposed in the mid-1990s by, e.g., Angrist et al. (1996). Under this framework, the population 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. This framework highlights the need for a monotonicity assumption, under which the "defiers" must be assumed not to exist in order to estimate a casual effect of interest. Under monotonicity (no defiers), IV estimates a local average treatment effect (LATE).[1]

[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

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.[2]

The main contribution of this paper is not in its theoretical econometric model of general first-stage heterogeneity, but rather in focusing on the implications of that heterogeneity for the small-sample bias of the estimator, and how researchers can take advantage of it. 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 IV estimate in expectation. This intuition about never-takers and always-takers extends to observations for which the instrument has a nonzero but small effect. Bias can be reduced by limiting the influence of these observations on estimation. Researchers already do this by, for example, selecting samples in which the instruments are likely to have an effect.

I derive the single-equation properties of two extremely simple estimators that directly model heterogeneity in the first stage. These estimators perform standard IV, except that the effect of the instrument is allowed to vary over groups, or is estimated at an individual level and then used as part of a sample weight.[3] As such, these new methods should be intuitive to users of regular IV, and can be implemented in any setting where linear IV is

(see, e.g., Heckman and Vytlacil, 2007). However, 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 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 Imbens and Angrist (1994). However, the simplified interpretation seems to have become common quickly, and can be found for example in Angrist and Imbens (1995). The ATE-among-compliers understanding appears to be common among applied researchers, and is often used in demonstrations of IV for student and researcher audiences (e.g. Imbens and Wooldridge, 2009; Wooldridge, 2010).

[3]To avoid introducing too many new terms in the paper, I refer to these estimators in the text as "SLATE estimators." However, I suggest "Magnified IV" as a general term, since these estimators magnify the impact of observations that respond strongly to the instrument.

performed. I additionally provide statistical packages to aid in the usage of these methods.[4]

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, (3) weaken the reliance on the monotonicity assumption in the group-interaction version of the estimator, and (4) give the researcher control over a tradeoff between bias and "localness" in the weighted version of the estimator. The weighted estimator also allows the ATE among compliers to be estimated, although this relies on large samples and very accurate estimates of individual first-stage treatment effects.

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, and so a more precisely-estimated SLATE may be preferable to a more-biased LATE. 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,[5] produces a SLATE rather than a LATE.

I explore the properties of the SLATE estimators relative to two stage least squares under different conditions including invalidity, heteroskedasticity, and violation of monotonicity, finding that the group-interaction version of the SLATE estimator performs well in the simulation settings explored, and also performs comparably to other weak-instrument methods despite being much simpler. The weighted SLATE estimator is not as successful.

The SLATE estimators rely on the ability to estimate variation in the first-stage treatment effect, and so 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

[4]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/")`.

[5]I will refrain from pointing fingers, but a literature search for "interact the instrument" produces many examples.

random selection ("GroupSearch"), and the Top-K τ-Path algorithm (TKTP) (Sampath and Verducci, 2013; Sampath et al., 2015, 2016; Bamattre et al., 2017). TKTP is intended to detect the sign of the relationship between the endogenous variable and instrument without the need to model that relationship with additional mediators. Neither GroupSearch nor TKTP are capable of precisely uncovering first-stage heterogeneity, but the SLATE estimators perform well regardless. The third approach is the causal forest (Athey and Imbens, 2016; Wager, 2018; Athey et al., 2019), which more precisely estimates heterogeneity in the treatment effect at the individual level by repeatedly partitioning the data using a set of high-dimensional controls, and 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 Angrist, Battistin, and Vuri (2017) and testing the ability to reproduce the full-sample estimate using small subsamples. In those subsamples, combining my estimators with causal forest reduces mean absolute error by about 22% in small ($N \approx 1,000$) subsamples, and by about 13% in larger ($N \approx 33,000$) subsamples.

II. INSTRUMENTAL VARIABLES WITH HETEROGENEOUS EFFECTS

II.i. 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 Imbens and Angrist (1994).

Consider a basic instrumental variables specification with one mean-zero endogenous variable x and one mean-zero exogenous variable z. Controls are not included, or they have been partialed out.

$$y_i = x_i\beta_i + \varepsilon_i \tag{1}$$

$$x_i = z_i\gamma_i + \nu_i \tag{2}$$

There is full heterogeneity in the effects of z_i on x_i (γ_i) and of x_i on z_i (β_i). Assume $E(z'\varepsilon) = E(z'\nu) = E(z'\gamma) = E(z'\beta) = E(x'\gamma) = E(x'\beta) = 0$ and $E(x'\varepsilon) = E(\nu'\varepsilon) \neq 0$, where a lack of an i subscript indicates a vector. The treatment effect varies with the effect of the instrument, so $E(\gamma\beta) \neq E(\gamma)E(\beta)$.

A standard IV estimator is calculated as:

$$\hat{\beta}^{IV} = \frac{N\widehat{Cov}(z,y)}{N\widehat{Cov}(z,x)} \tag{3}$$

where N is the sample size.

$$\begin{aligned} N\widehat{Cov}(z,y) = \sum_i z_i y_i = \sum_i z_i(x_i\beta_i + \varepsilon_i) = \sum_i (z_i x_i \beta_i + z_i\varepsilon_i) \\ = \sum_i (z_i(z_i\gamma_i + \nu_i)\beta_i + z_i\varepsilon_i) = \sum_i (z_i^2\gamma_i\beta_i + z_i\nu_i\beta_i + z_i\varepsilon_i) \end{aligned} \tag{4}$$

$$N\widehat{Cov}(z,x) = \sum_i z_i x_i = \sum_i (z_i^2\gamma_i + z_i\nu_i) \tag{5}$$

In expectation, since $E(z'\varepsilon) = E(z'\nu) = Cov(z,\gamma) = Cov(z,\beta) = 0$, this becomes

$$E(\hat{\beta}^{IV}) = \frac{E(\gamma'\beta)}{E(\gamma)} \tag{6}$$

The expected value of the IV estimator is a weighted average of the β_is, where the weights

are γ_i. This does not match the common interpretation among applied researchers that the LATE is the ATE among compliers. The common interpretation only holds if γ_i is limited to only 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:

$$\frac{\sum_i z_i \nu_i \beta_i}{\sum_i z_i x_i} + \frac{\sum_i z_i \varepsilon_i}{\sum_i z_i x_i} \tag{7}$$

The second of these is well known from any IV derivation. The first is present because of the assumption that $E(\gamma\beta) \neq E(\gamma)E(\beta)$, and so $E(x'\beta) \neq E(x)E(\beta)$, preventing a term from canceling out as normal.

This basic derivation isolates several points about IV, most of which are well-known:

1. If γ_i takes both positive and negative signs (monotonicity does not hold), standard IV generally does not estimate a parameter of interest.

2. If γ_i takes a range of values, the IV estimand is a weighted average of treatment effects where the weights are γ_i.

3. In finite samples, IV is biased.

4. The size of the IV bias is based on $\sum_i z_i \nu_i \beta_i$ and $\sum_i z_i \varepsilon_i$, and is smaller the stronger the relationship is between z_i and x_i.

In addition, this makes clear that observations with γ_i close to 0 do not have an effect on the expected value of the IV estimand. However, in a finite sample, the IV bias term has $\sum_i z_i x_i = \sum_i (z_i^2 \gamma_i + z_i \nu_i)$ in the denominator. The addition of a single observation $N+1$ to the sample where $\gamma_{N+1} = 0$ will not change the expected estimate at all, but will increase the numerator of the bias by $z_{N+1}\nu_{N+1}\beta_{N+1} + z_{N+1}\varepsilon_{N+1}$ and the denominator by $z_{N+1}\nu_{N+1}$. Unless $\beta_{N+1} = 0$ and $Var(\varepsilon) \geq Var(\nu)$, or some relaxed combination of the two,

this introduces more noise into the numerator than the denominator, increasing the extent to which variation in the estimate is driven by bias rather than sampling variation.

So, despite not affecting the identified parameter of interest or the expected value of the IV estimator, these observations do introduce additional noise to the estimator, make the instrument weaker, and worsen the small-sample properties of the IV estimator.

One potential means of improving the small-sample properties of the IV estimator, then, is to find and remove or downweight observations with small values of γ_i, which should increase the absolute value of $Cov(z_i, x_i)$ and reduce bias.

II.ii. 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 $g_i \in \{1, ..., G\}$, and the coefficient on the instrument is allowed to vary over those groups. Controls and group fixed effects have been partialed out in both the first and second stages. The true model is the same as in the previous section, but the estimation model becomes:

$$y_i = x_i\beta_i + \varepsilon_i \tag{8}$$

$$x_i = z_i \sum_g \gamma_g I_{gi} + \nu_i \tag{9}$$

where I_{gi} is an indicator function equal to 1 if $g_i = g$. Estimating this model by 2SLS, 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.

$$\hat{x}_i = z_i \sum_g \hat{\gamma}_g I_{gi} = z_i \sum_g \frac{Cov(x_i, z_i|I_{gi})}{Var(z_i|I_{gi}))} I_{gi} \tag{10}$$

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

$$\hat{\beta}^{2SLS} = \frac{N(\widehat{Cov}(\hat{x}, y))}{N(\widehat{Var}(\hat{x}))} \tag{11}$$

The numerator and denominator can be expanded as

$$\begin{aligned} N(\widehat{Cov}(\hat{x}, y)) = \sum_i \hat{x}_i y_i &= \sum_i z_i y_i \sum_g \hat{\gamma}_g I_{gi} \\ &= \sum_i (z_i^2 \gamma_i \beta_i + z_i \nu_i \beta_i + z_i \varepsilon_i) \sum_g \hat{\gamma}_g I_{gi} \\ &= \sum_g \hat{\gamma}_g \left(\sum_i (z_i^2 \gamma_i \beta_i + z_i \nu_i \beta_i + z_i \varepsilon_i) I_{gi} \right) \end{aligned} \tag{12}$$

$$N(\widehat{Var}(\hat{x})) = \sum_i \hat{x}_i^2 = \sum_i \left(z_i \sum_g \hat{\gamma}_g I_{gi} \right)^2 = \sum_g \hat{\gamma}_g^2 \left(\sum_i z_i^2 I_{gi} \right) \tag{13}$$

In expectation, $E(\hat{\gamma}_g) = \frac{1}{N_g} \sum_i \gamma_i I_{gi} \equiv \gamma_g$ and $E(z_i \nu_i) = E(z_i \varepsilon_i) = 0$. As a result, 2SLS identifies

$$E(\hat{\beta}^{2SLS}) = \frac{E(\sum_i \beta_i \gamma_i \sum_g \gamma_g I_{gi})}{E(\sum_g \gamma_g^2 N_g)} \tag{14}$$

where $N_g = \sum_i I_{gi}$ is the number of individuals in group g. This is a weighted average of the β_is, where the weights are $\gamma_g \gamma_i$ for the associated γ_g. This narrows the monotonicity assumption to instead be monotonicity-within-group, i.e. that γ_g and γ_i always have the same sign $\forall\ g_i = g$ so weights are positive. As $Var(\gamma_g)$ approaches zero among groups with non-zero γ_g, as might occur if there were no differences between groups, or if the treatment effect were either zero or a constant c (as in the basic defiers-compliers framework), the closer one of the γ_g terms comes to canceling out, returning to the γ_i LATE weights of the previous section.

If γ_i is constant within group, this simplifies to

$$E(\hat{\beta}^{2SLS}) = \frac{E(\sum_i \beta_i \sum_g \gamma_g^2 I_{gi})}{E(\sum_g \gamma_g^2 N_g)} \tag{15}$$

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 this class of estimates as being Super-Local Average Treatment Effects (SLATE).

In a finite sample, the bias term is

$$\frac{\sum_g \hat{\gamma}_g \left(\sum_i z_i(\nu_i\beta_i + \varepsilon_i) I_{gi}\right)}{\sum_g \hat{\gamma}_g^2 \left(\sum_i z_i^2 I_{gi}\right)} \tag{16}$$

Compared to the bias term in the previous section, 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:

$$\frac{\sum_g (\hat{\gamma}_g - \hat{\gamma}) \left(\sum_i z_i(\nu_i\beta_i + \varepsilon_i) I_{gi}\right) + \hat{\gamma} \sum_i z_i(\nu_i\beta_i + \varepsilon_i)}{\sum_g (\hat{\gamma}_g^2 - \hat{\gamma}^2) \left(\sum_i z_i^2 I_{gi}\right) + \hat{\gamma}^2 \sum_i z_i^2} \tag{17}$$

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

$$\frac{\sum_g E\left(\hat{\gamma}_g^2 \left(\sum_i z_i^2 (\nu_i\beta_i + \varepsilon_i)^2 I_{gi}\right)\right)}{\sum_g E\left(\hat{\gamma}_g^4 \left(\sum_i z_i^4 I_{gi}\right)\right)} \tag{18}$$

By the BLUE properties of OLS, estimating the first stage separately by group will necessarily increase $\widehat{Var}(\hat{x})$.[6] So, taking $\hat{\gamma}_g = \gamma_g$ and assuming that the $(\nu_i\beta_i + \varepsilon_i)^2$ 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

[6] Recall that between-group differences have already been partialled out from both x and z, so Simpson's paradox does not apply here.

also always the possibility that in a given finite sample, $\hat{\gamma}_g$ may be related to $z_i^2(\nu_i\beta_i + \varepsilon_i)^2$, increasing bias relative to regular IV.

Compared to the previous section, allowing the effect of the instrument to vary over groups serves two purposes: it generally reduces bias, and also it increases the weight of the estimator on the β_is associated with high γ_i values. In other words, it increases the "localness" of the estimate. This implies, in instrumental variables, a tradeoff between bias and localness.

II.iii. 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. The weighted IV estimate β^{WIV} is

$$\hat{\beta}^{WIV} = \frac{N\widehat{Cov}(Wz, Wy)}{N\widehat{Cov}(Wz, Wx)} \tag{19}$$

Where W is a diagonal matrix with $w = \{w_1, w_2, ...\}$ on the diagonal. Assume that weights are chosen such that $Cov(w, z) = 0$.

Following the same derivations as in the previous section,

$$E(\hat{\beta}^{WIV}) = \frac{E((WW\gamma)'\beta)}{E(WW\gamma)} \tag{20}$$

In other words, weighted IV estimates a weighted average of the β_is, where the weights are $w_i^2\gamma_i$. In finite samples, the bias terms are

$$\frac{\sum_i w_i^2 z_i \nu_i \beta_i}{\sum_i w_i^2 z_i x_i} + \frac{\sum_i w_i^2 z_i \varepsilon_i}{\sum_i w_i^2 z_i x_i} \tag{21}$$

Assume also that γ is known. If weights are chosen such that $Cov(w, z) = 0$ but $Cov(w, \gamma) > 0$, this will identify a SLATE and reduce variation in the bias term.

There are many such weights that could fulfill the role of being independent of z but related to γ. However, there are three weighting functions that may be of particular interest.

The first is an indicator function $w_i = I(\gamma \neq 0)$. This effectively drops all observations with $\gamma_i = 0$ from the sample, which will strengthen the instrument and reduce variance in the bias term. This will also 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.

The second is $w_i = (F_{\gamma_i})^p$ for some $p \neq 0$,[7] 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 setting $w_i = |\gamma_i|^{4p}$ 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 the amount of bias: an increase in p will usually reduce bias (proof to follow), but will make the estimate more heavily weight observations with large γ_i values, increasing the "localness" of the estimate. In effect, there is a bias-localness trade-off, and the researcher has some control over that trade-off.

Setting $p = 1/4$ is a natural choice. 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. If γ_i is constant within group and monotonicity holds, the two weights are identical.

Another natural choice for p is $p = -1/4$ (and $w_i = 0 \ \forall \ \gamma_i = 0$), even though it worsens small-sample bias. When $p = -1/4$, if the sign of γ_i is constant (no defiers), then in the single-instrument setting the proposed estimator identifies a weighted average of the β_is with 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.

Given these possible weighting schemes, it is important to determine the impact of p on

[7] If $p < 0$, weighting should set $w_i = 0$ if $F_{\gamma_i} = 0$ to avoid dividing by 0.

IV bias. With $w_i = (F_{\gamma_i})^p$ weighting, the bias in a single-instrument setting is:

$$bias = \left(\frac{\sum_i (f\gamma_i^2)^p z_i(\nu_i \beta_i + \varepsilon_i)}{\sum_i (f\gamma_i^2)^p z_i x_i} \right) \equiv \zeta \tag{22}$$

where $f = (N-k)Var(z)/Var(M_z x)$ and M_z is the z elimination matrix.

$$\frac{\partial \zeta}{\partial p} = \frac{\sum_{i|\gamma_i \neq 0} \log(f\gamma_i^2)(f\gamma_i^2)^p z_i(\nu_i \beta_i + \varepsilon_i)}{\sum_i (f\gamma_i^2)^p z_i x_i} - \zeta \frac{\sum_{i|\gamma_i \neq 0} \log(f\gamma_i^2)(f\gamma_i^2)^p z_i x_i}{\sum_i (f\gamma_i^2)^p z_i x_i} \tag{23}$$

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.[8] If Equation 23 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.

Does the second term dominate Equation 23? 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, ν_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.[9] As the sample size grows, the chance that

[8] 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.

[9] Alternately, consider the variance of the bias, ζ^2. Under i.i.d., all cross-product terms drop out, and the structure of $\partial \zeta^2 / \partial p$ is very similar to Equation 23; when the variance of bias is large, the second term dominates and p reduces variance in bias. When variance of bias is small, noise in the first term dominates.

a reduction in p might reduce bias increases.

These results are dependent upon using known values of γ_i. The performance of the weighting estimator when γ_i is poorly estimated is not as certain. 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 more dependent on identifying $\hat{\gamma}$ precisely. 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 under which the weighted estimator improves performance hold.

III. FEASIBLE ESTIMATORS FOR SIMULATION

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. If the effect of z_i on x_i is expected to vary over a set of covariates v_i, then an interaction between z_i and v_i can be included in the model, or γ_i can be allowed to vary over v_i in a multilevel model (Raudenbush and Bryk, 2002), or a number of other methods, including recent developments in machine learning for modeling heterogenous treatment effects like causal forest (Athey and Imbens, 2016; Wager, 2018; Athey et al., 2019). Any such approach would allow the group-based method in Section II.ii to be performed. Alternately, any method that models γ_i directly can be used to follow the weighting method in Section II.iii, or to combine it with the group-based method. Since the SLATE estimators can include controls for v_i in both stages, these approaches do not require a validity assumption for v_i.

This section, and the following simulation, will focus on two methods for estimating first-

stage heterogeneity that do not require any additional information about variation in γ_i, instead trying to identify groups g over which γ_i varies from the data itself. 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 IV.iii. Despite this, the SLATE estimators still perform well.

As demonstrated in Sections II.ii and II.iii, improved performance relies on the ability to select groups over which γ_i actually varies, or to estimate γ_i accurately so that the weights w_i can have a positive relationship with γ_i. Approaches that use all available information such as covariates are likely to improve performance further. I will use causal forest to model first-stage heterogeneity using controls in Section V. The simulation results should be understood as the ability of the estimators to improve performance even without the benefit of additional information about heterogeneity.

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

GroupSearch, with enough iterations, should be able to identify groups within the data for which there is between-group variation in $\hat{\gamma}_g$. In simulation, I attempt 100 different randomly-selected groupings for each sample.

There is the potential concern that GroupSearch will introduce bias either via overfitting or by inducing some correlation with second-stage error term ν and invalidating the instrument. However, these are unlikely to be major issues.

The overfitting concern is valid, but only for the first stage: the estimate of the relationship between x and z will be overfitted. 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., Belloni et al. 2014). Overfitting is acceptable.

The concern that GroupSearch might invalidate the instrument would require that z be invalid in the first place. At least in the Section II.ii methods, the grouping structure is to be partialed out or controlled for, and so any relationship between the grouping structure itself and ν is accounted for by the method. 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 ν 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 already be invalid. As long as z_i is truly valid, this should not be possible.

The second method is the Top-K τ-Path search, or TKTP (Sampath and Verducci, 2013; Sampath et al., 2015, 2016; Bamattre et al., 2017). Given two variables (x and z in our 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. For any two observations, x_i, z_i, x_j, and z_j, that pair is concordant if $x_i > x_j$ and $z_i > z_j$, or if $x_i < x_j$ and $z_i < z_j$, and discordant otherwise. Kendall's τ is the proportion of pairs that are concordant. A higher τ indicates a stronger positive relationship between x and z. TKTP creates a τ-path by arranging the observations in order such that, if $\tau(i)$ is τ calculated using the first i observations in that order, τ 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, such that the observations $\{1, 2, ..., j\}$ in the τ-path are considered to have a positive relationship, and $\{j + 1, ..., n\}$ are not.

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.[10]

[10]Because there is some randomness injected in the algorithm, it is possible that the same observation

Since it specifically tests for subgroups with positive and negative associations separately, TKTP seems ideal in cases where there may be an unmeasured defier subgroup. A downside of TKTP is that, under current implementations, it is computationally slow, and may not be usable for very large data sets.

IV. SIMULATION

I test the properties of the proposed estimators under simulated-data settings, beginning with a setting where all IV assumptions are satisfied, and then in subsections evaluating alternate data generating processes (DGP), some of which contain the violation of standard assumptions.

Data simulation centers around the data-generating process

$$y_i = x_i\beta_i + 2w_i + \varepsilon_i \tag{24}$$

$$x_i = z_i\gamma_i + w_i + \nu_i \tag{25}$$

$$z_i, w_i, \varepsilon_i, \nu_i \sim N(0, 1) \tag{26}$$

where w_i represents an unobserved confounding factor, and is not controlled for in analysis. 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 $1,600$. 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. Then I use those groups to estimate γ_gs to use for weights

may end up in both groups, in which case it is assigned to neither.

with $p = 1/4$ for the weighting and combined group/weight versions of the SLATE estimator. I compare estimates to the true LATE and SLATE given the formulae in Sections II.i and II.ii.

IV.i. BASIC SIMULATION

Here I present results following the DGP in Section IV. I present feasible results, taking as known only the number of underlying groups for use with GroupSearch. I implement both GroupSearch and TKTP for feasible estimation. TKTP is not implemented for sample sizes above $1,600$ due to computational limitations.[11]

Figure 1 shows performance using feasible estimation. The GroupSearch-selected groups improve upon 2SLS by about 50% at the $N = 1,600$ point. Adding weights on top of the group modeling does not change performance. Top-K τ-Path underperforms relative to GroupSearch, even though it uses a more rigorous approach to identifying treatment effect variation.

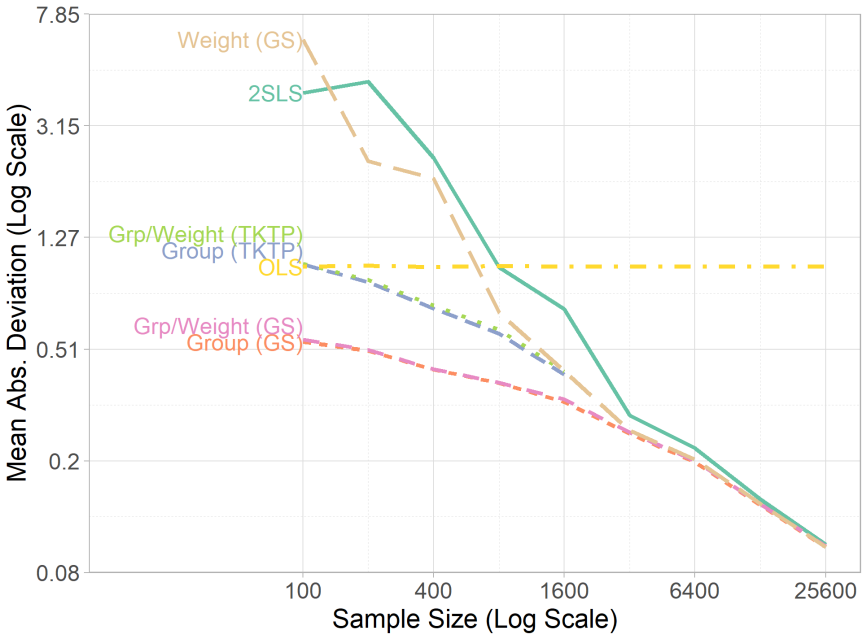
In general, the proposed group-based estimator considerably outperforms 2SLS at smaller sample sizes, and is very similar to 2SLS at large sample sizes. The weighted versions do not perform as well.

Bootstrap standard errors are higher than for OLS, as shown in Figure 2. But they are in most forms better than 2SLS at small sample sizes, and similar at large sample sizes, although the proposed estimators do not outperform 2SLS by as large a margin on standard error as they do on deviation.

Performance is similar using $\gamma_i \sim U[0, 1/4.5]$, which is chosen to retain treatment effect averages with the original DGP. In this and every other simulation using a continuous distri-

[11] The slow part of TKTP is the Backwards Conditional Search (BCS) part of their algorithm. In this paper I use the FastBCS R implementation in Caloiaro (2019) from one of the original authors of Sampath et al. (2015), and combine FastBCS with my own code for the rest of the algorithm. Other non-R implementations are faster, so the use of TKTP with larger samples is feasible, especially if it only needs to be run once rather than 1,000 times.

Figure 1: Performance Using Feasible Estimation



Deviation is relative to parameter identified in expectation. At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of four groups from the best of 100 random groupings generated. Top-K τ-Path (TKTP) uses TKTP to identify groups in which z and x have positive, negative, or null relationships, respectively. TKTP is only run for smaller samples due to computational limitations. For Weight variants, first stage coefficients are estimated using groups, and then those coefficients are used to generate weights. See Section IV for data-generating process.

Figure 2: Standard Error Using Feasible Estimation

At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of four groups from the best of 100 random groupings generated. Top-K τ-Path (TKTP) uses TKTP to identify groups in which z and x have positive, negative, or null relationships, respectively. TKTP is only run for smaller samples due to computational limitations. For Weight variants, first stage coefficients are estimated using groups, and then those coefficients are used to generate weights. See Section IV for data-generating process.

bution 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. See Appendix Appendix A Figure A.13.

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

IV.ii. INVALIDITY

IV relies on a validity assumption for consistency. It is possible that the nature of the proposed estimators, which attempt to maximize the influence of the instruments, may make the estimate more sensitive to validity violations, as described in Section III. To test for the impact of minor violations of validity, I generate z_i as

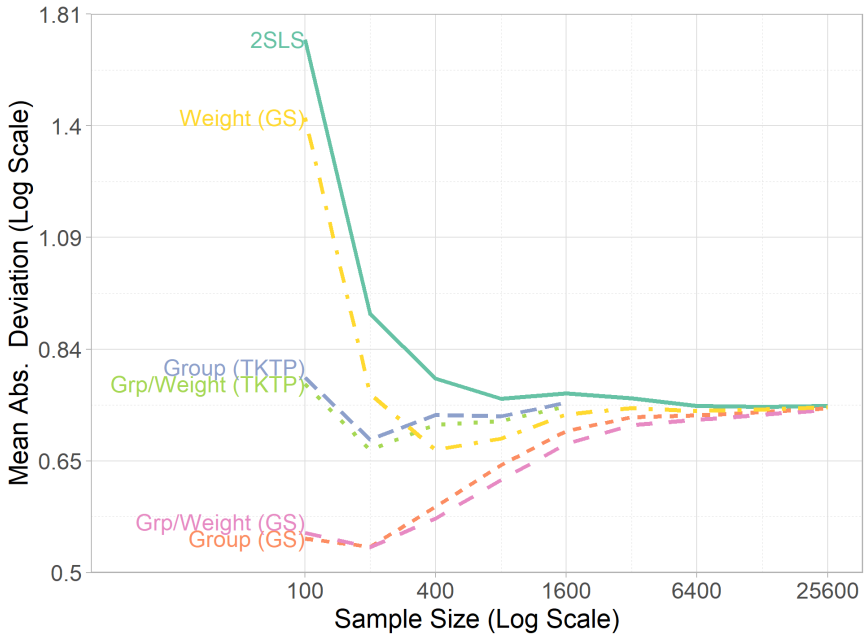
$$z_i = .2w_i + \zeta_i; \zeta_i \sim N(0, 1) \tag{27}$$

The results of this simulation can be seen in Figure 3. Under this violation, all IV variants converge to a higher level of deviation than in previous sections, which is to be expected since the estimator is inconsistent. But at each sample size, the proposed estimators continue to outperform 2SLS. Under the violation of validity, there is less deviation for the proposed estimators at small sample sizes at large sample sizes.[12]

In addition to standard violations of validity, the proposed estimators introduce the possibility that γ_i will be related to the second-stage error term. If this occurs, then using

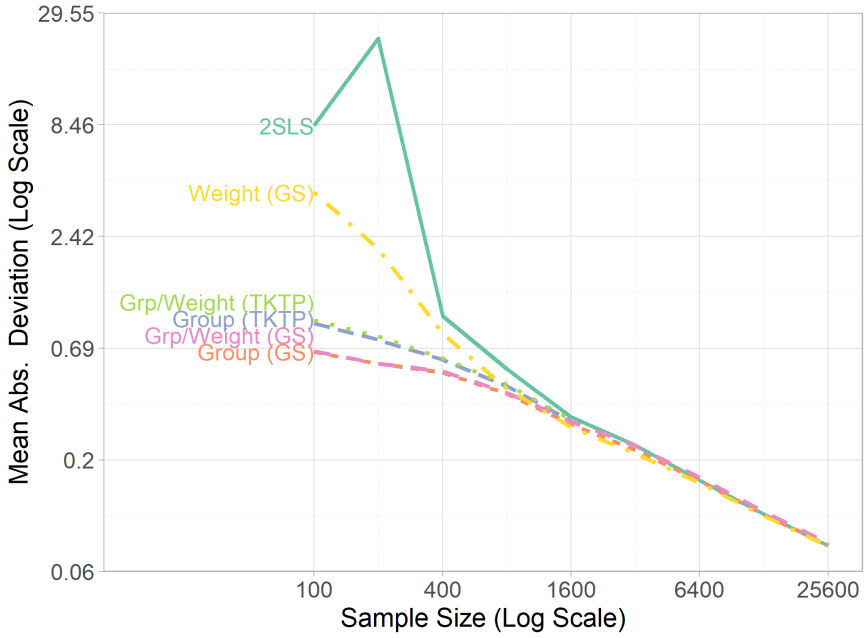
[12]Performance actually worsens in larger samples here for GroupSearch. This is because, in smaller samples, it is likely that some of the small-sample variation between groups picked up by GroupSearch is unrelated to the true underlying invalidity. As samples increase, GroupSearch picks up this invalid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

Figure 3: Performance With Validity Violation



Deviation is relative to parameter identified in expectation. At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of four groups from the best of 100 random groupings generated. Top-K τ-Path (TKTP) uses TKTP to identify groups in which z and x have positive, negative, or null relationships, respectively. TKTP is only run for smaller samples due to computational limitations. For Weight variants, first stage coefficients are estimated using groups, and then those coefficients are used to generate weights. See Section IV.ii for data-generating process.

Figure 4: Performance with Validity Violation in γ_i



Deviation is relative to parameter identified in expectation. At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of four groups from the best of 100 random groupings generated. Top-K τ-Path (TKTP) uses TKTP to identify groups in which z and x have positive, negative, or null relationships, respectively. TKTP is only run for smaller samples due to computational limitations. For Weight variants, first stage coefficients are estimated using groups, and then those coefficients are used to generate weights. See Section IV.ii for data-generating process.

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

$$\gamma_i = \phi_i + .05(w_i - \min(w_i))/\max(w_i)) \tag{28}$$

where $\phi_i \sim U[0, 1/4.5]$. The results of this simulation are in Figure 4. Under this violation, the proposed estimators are still no worse than 2SLS, and are considerably better for very small samples, but the proposed estimators and 2SLS reach similar levels of mean absolute deviation at smaller sample sizes than in Figure 1, around 1,600 observations rather than 6,400.

IV.iii. MONOTONICITY

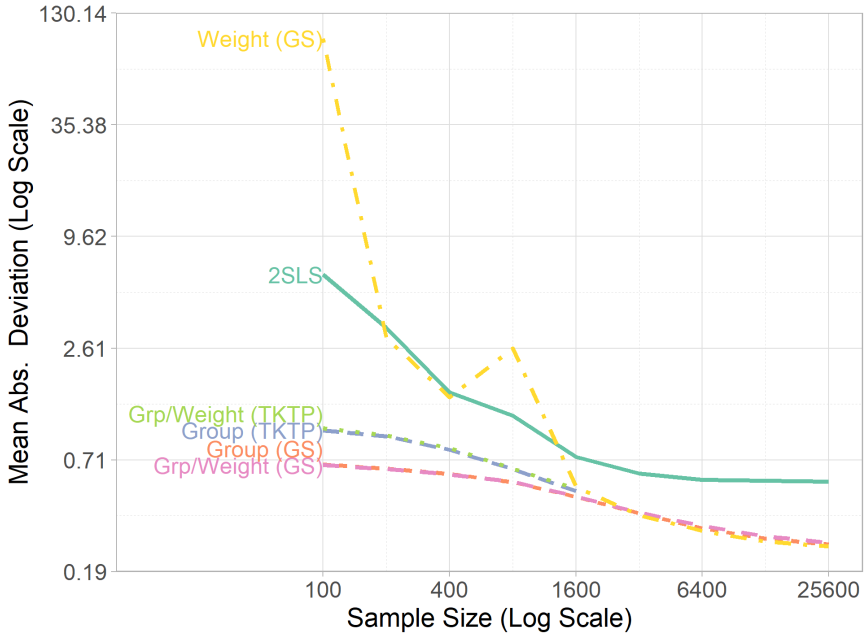
In cases where monotonicity is violated, the IV estimand is not of particular interest, as it contains negative weights on some treatment effects. This is true for 2SLS, and is also true for the proposed estimators unless the subsample of "defiers" can be identified for each instrument and the effect of the instrument is allowed to be different for that group.

If the underlying group structure is known, then the group-based estimator is proven in previous sections to identify the SLATE even under violations of between-group monotonicity. But this does not ensure that the feasible estimators can identify the group structure. I repeat the DGP from Section IV except that $\gamma_i \sim U[-1/9, 3/9]$. I then perform GroupSearch with four groups and also TKTP.

The ability of both methods to identify the defier groups is underwhelming. I perform a chi-square test to look for a relationship between the TKTP-identified groups and the groups with true negative or positive effects. The p-value from this test is around .05 at all sample sizes evaluated. TKTP tended to overassign observations to the "No relationship" group - "No relationship" was the modal group assigned by TKTP to true-negative observations across all sample sizes. Excluding "No relationship" so the only options are positive and negative, 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 tested. 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. Across all 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.

However, even though many observations are misclassified, classification is a considerable improvement over no classification, especially given that 2SLS performance is worse here than in other simulations. Figure 5 shows the results of the simulation, in which the proposed estimators perform better under violations of monotonicity than 2SLS does, and the improvement is to a greater degree than in Figure 1. At the $N = 1,600$ point, for

Figure 5: Performance under Violation of Monotonicity



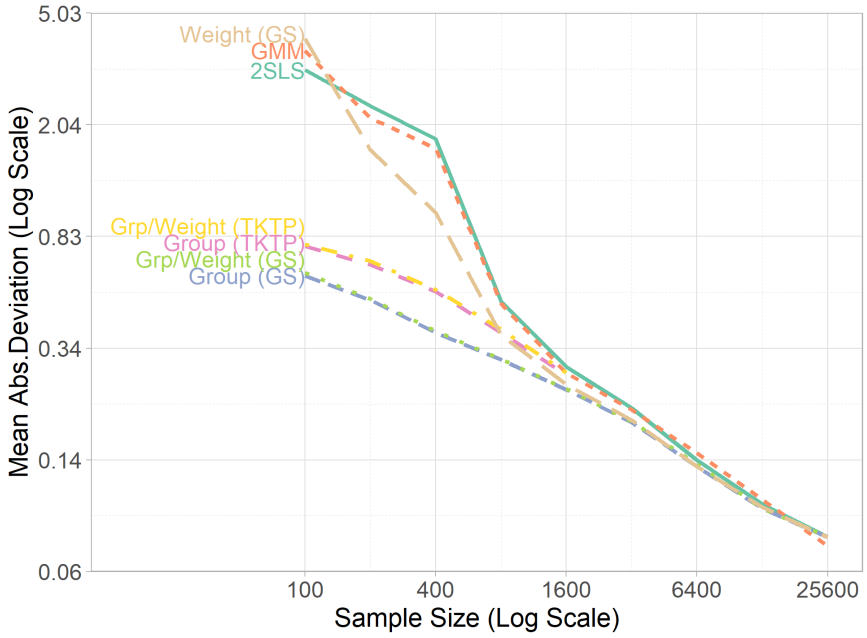
Deviation is relative to LATE or SLATE, as appropriate, with all-positive weights. At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of four groups from the best of 100 random groupings generated. Top-K τ-Path (TKTP) uses TKTP to identify groups in which z and x have positive, negative, or null relationships, respectively. TKTP is only run for smaller samples due to computational limitations. For Weight variants, first stage coefficients are estimated using groups, and then those coefficients are used to generate weights. See Section IV.iii for data-generating process.

example, the GroupSearch approach reduced mean absolute deviation in Figure 1 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.

IV.iv. CLUSTERING

As demonstrated in Young (2018), 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 Young (2018), 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

Figure 6: Performance under Clustering



Deviation is relative to parameter identified in expectation. At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of four groups from the best of 100 random groupings generated. Top-K τ-Path (TKTP) uses TKTP to identify groups in which z and x have positive, negative, or null relationships, respectively. TKTP is only run for smaller samples due to computational limitations. For Weight variants, first stage coefficients are estimated using groups, and then those coefficients are used to generate weights. See Section IV.iv for data-generating process.

$$x_i = z_i\gamma_i + \lambda_c(\eta_c + w_i + nu_i)\sqrt{2} \tag{29}$$

$$y_i = x_i\beta_i + \lambda_c(\eta_c + 2w_i + \varepsilon_i)/\sqrt{2} \tag{30}$$

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 6 shows the results of the simulation. The proposed 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 proposed estimators may be harmed more in relative terms by clustering than 2SLS is. However, the proposed estimators still outperform 2SLS in this clustered setting.

IV.v. OTHER WEAK-INSTRUMENT METHODS

The proposed estimators are not the only existing approach to reducing small-sample bias. Previously existing alternatives include variations of Limited-Information Maximum Likelihood (LIML), and the Jackknife Instrumental Variables Estimator (JIVE). I compare the performance of the proposed estimators to JIVE and to the Fuller (1977) implementation of LIML, using the DGP from Section IV.[13] LIML is a $k-$class estimator known to be biased, but Fuller (1977) suggests an adjustment parameter α for k which he suggests be set to $\alpha = 1$ for unbiasedness or $\alpha = 4$ for minimum mean squared error. I run both, as "Fuller (1)" and "Fuller (4)".

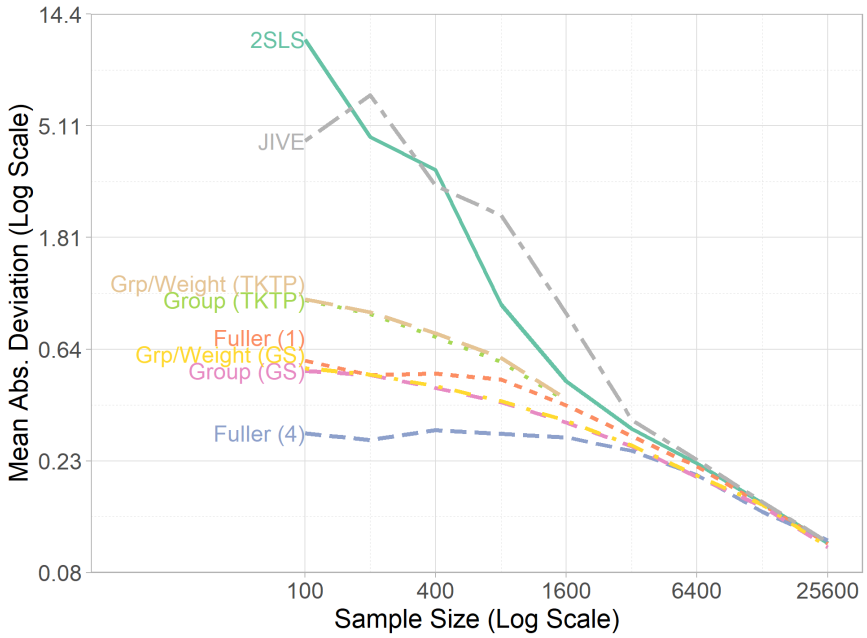
Figure 7 compares all of these estimators. The performance of JIVE is fairly weak in the given setting, not outperforming even 2SLS. The Fuller (1) implementation of LIML, however, has similar performance to the proposed estimators, and outperforms the Top-K τ-Path variant, but is modestly outperformed by the grouping estimator in terms of deviation. Fuller (4) outperforms all SLATE estimators in mean absolute bias. However, it does return a biased result (Fuller, 1977), 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, although there are heteroskedasticity-robust variants such as Hausman et al. (2012). There are many other small-sample robust estimators that could be tried.

IV.vi. NUMBER OF GROUPS

The final two simulation subsections, instead of testing performance under violated assumptions or in comparison to other methods, check the performance of the SLATE estimator under different settings - first, testing the impact of the choice of the number of groups to

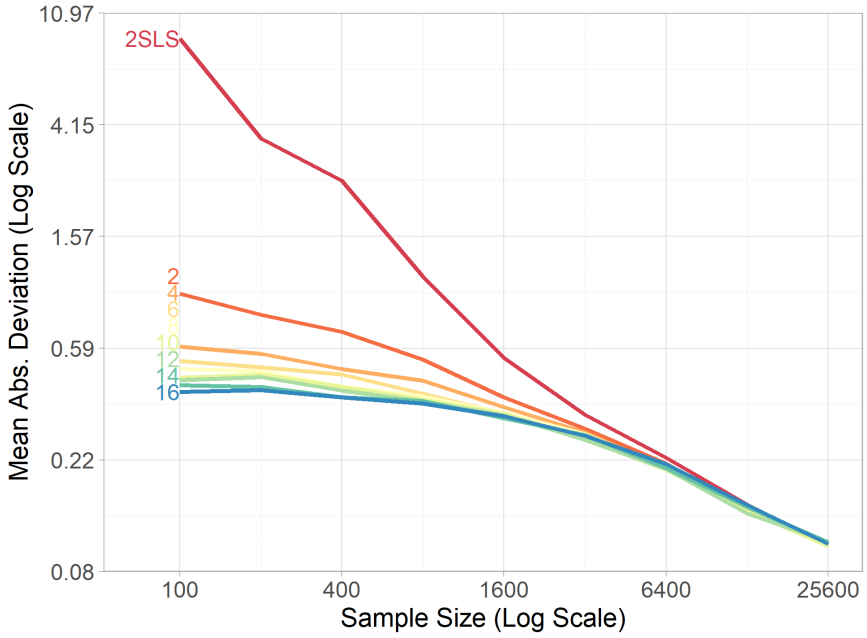
[13] Fuller and JIVE use the implementations in Jiang et al. (2017) and Ginestet (2016), respectively.

Figure 7: Comparison of Proposed Estimators to Other Weak-Instrument Methods



Deviation is relative to parameter identified in expectation. At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of four groups from the best of 100 random groupings generated. Top-K τ-Path (TKTP) uses TKTP to identify groups in which z and x have positive, negative, or null relationships, respectively. TKTP is only run for smaller samples due to computational limitations. For Weight variants, first stage coefficients are estimated using groups, and then those coefficients are used to generate weights. See Section IV.v for data-generating process.

Figure 8: GroupSearch Using Different Numbers of Groups



Deviation is relative to parameter identified in expectation. At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of the specified number of groups from the best of 100 random groupings generated. See Section IV.vi for data-generating process.

model, and second, attempting to recover the ATE among compliers, as described in Section II.iii.

While the TKTP feasible estimator will naturally produce 1-3 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 Section V. Cross validation, a standard tool for selecting parameters, does not make much sense for GroupSearch where 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 $\{2, 4, ..., 16\}$. 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.

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 $Var(\hat{\gamma}_g)$ 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 could be improved further with the use of more groups.

IV.vii. RECOVERING THE ATE AMONG COMPLIERS

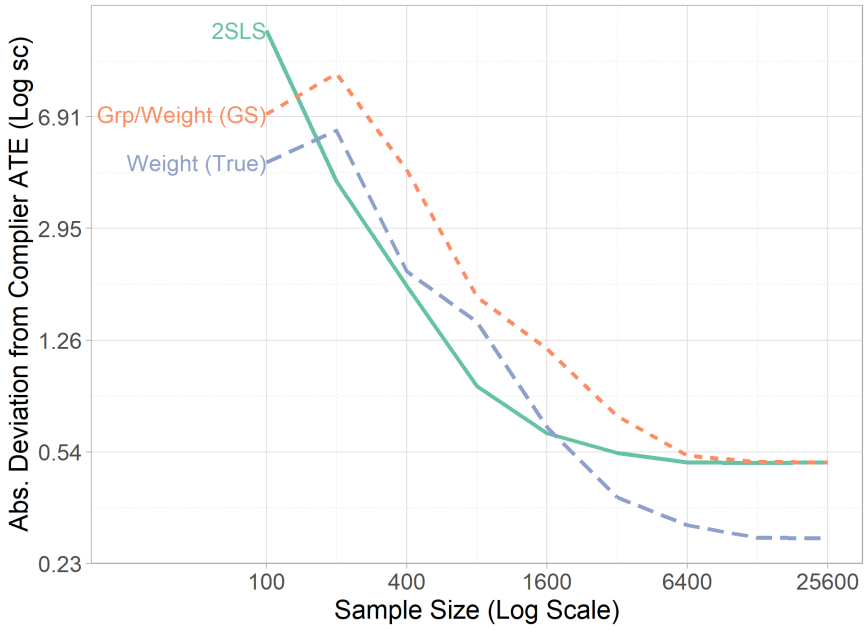
Section II.i shows that, unless the effect of the instrument takes only two values, one of which is 0 (as in the canonical LATE description), then the LATE identified by IV is not equivalent to "the ATE among compliers." As discussed in Section II.iii, however, the ATE among compliers can be recovered if there are no defiers, and the proposed weighting estimator uses a weighting scheme where $w_i = 0 \; \forall \; \gamma_i = 0$ and $w_i = (F_{\gamma_i})^{-1/4}$ otherwise.

Here I use the original DGP from section IV of $\beta = \{1, 2, 3, 4\}$ and $\gamma = \{0, .075, .15, .223\}$ and apply the ATE-recovering weighting scheme. 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: 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 $\hat{\gamma}_i \leq 0$.

Figure 9 shows the results. As demonstrated, the weighting 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. That this only works with large samples, even with true γ_is, makes sense. 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_is can be made.

Figure 9: Deviation from ATE Among Compliers



At each sample size, 1,000 random samples are drawn. Weight estimates use a GroupSearch (GS) grouping of four groups from the best of 100 random groupings generated. The first stage coefficients are estimated using those groups, and then those coefficients are used to generate weights. See Section IV.vii for data-generating process.

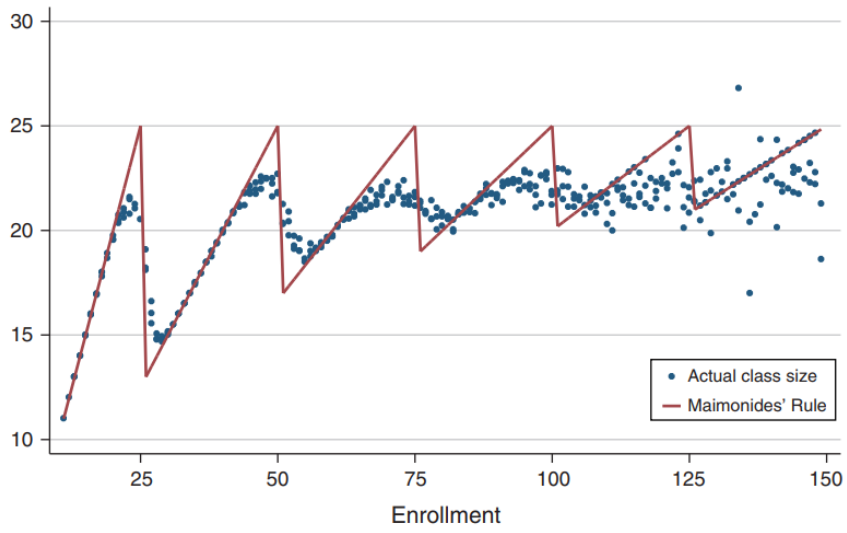
V. APPLICATION

In this section I demonstrate the real-world applicability of the proposed grouping estimator by replicating Angrist, Battistin, and Vuri (2017) (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 due to the fact that 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 (Angrist and Lavy, 1999).

ABV offers a useful setting for replication in this paper. First, data and replication code is freely available.[14] Second, ABV allows me to demonstrate the use of the proposed estimator in a multiple-instrument setting. Third, the sample is large enough that I can

[14]See https://www.aeaweb.org/articles?id=10.1257/app.20160267

Figure 10: Grade 5 Pre-Reform Enrollment and Class Size (ABV Figure 2b)



demonstrate the small-sample properties of the estimator by selecting subsamples of different sizes. Fourth, as will be shown, the instrument in ABV is very strong, and so replication will demonstrate that the usefulness of the proposed estimators is not limited to cases of weak instruments.

Fifth, this is a setting in which the effect of the instrument should vary over the sample. While monotonicity seems likely to hold, adherence to the Maimonides rule is not perfect. Compliance can be graphically shown to vary with enrollment, and presumably varies by other factors as well. Figure 10, copied from ABV Figure 2b, 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* × *grade* level. A long list of controls are included.[15]

[15]Controls include percent female, percent immigrant, father's education, mother's employment status, school enrollment by grade (and squared), distance from class-size-minimum threshold (and distance interacted with enrollment and enrollment squared), survey year, grade, grade enrollment at institution, region (and region interacted with grade enrollment), and percent students with missing (respectively) gender, origin, mother's education, and mother's occupation.

Analysis is performed separately by region.

Table 1 Panel A replicates ABV Table 6. Then, in Panel B, I use GroupSearch as described in Section III to perform the SLATE grouped estimator. Separately for monitored and non-monitored contexts, I randomly assign five different group identities 100 times, interact group identity with the instrument, and select the grouping that provides the highest first-stage F-statistic. 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 IV showed that the weighting estimator does not perform well in idealized settings.

In Panel C, I use causal forests (Athey and Imbens, 2016; Wager, 2018; Athey et al., 2019) to estimate a first-stage effect for each individual, allowing the effect to vary with all covariates. Causal forest is an extension of random forest methods. In a random forest, trees are built by iteratively splitting the sample to reduce prediction error within each split. Causal forests take a similar approach, but instead of reducing prediction error, they maximize the difference between splits in the estimated treatment effects. I use default "honest" causal forest estimations from the R package grf (Athey et al., 2019). Because overfitting is not a concern, as discussed in Section III, I generate individual treatment effect estimates for the full sample rather than using a holdout. The identifying assumptions necessary to treat causal forest estimates as causal are satisfied in the first stage by the standard validity assumption of IV.

Using the individual-level treatment effect estimates from the causal forest, I divide the sample into quintiles based on their estimated effect to form five groups. I use these groups to implement the SLATE grouped estimator. The causal forest is performed separately for monitored and non-monitored settings, generating different groupings for each.

Regular IV and the GroupSearch 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 proposed 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, but uses many more instruments. As a result, with five times as many instruments, the first-stage F statistics are slightly more than 1/5 as large.

The lower F-statistic does not translate into worse performance, however. I focus on the All Italy math score results from Table 1, and estimate the 2SLS and SLATE estimators by cluster bootstrap (where a number of clusters equal to the original number of clusters $C = 28,546$ are selected with replacement), producing 1,000 cluster bootstrap samples and performing 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 Young (2018), any of the estimation methods is particularly sensitive to the removal of certain clusters, this will be apparent in the results.

Then, I repeat the cluster bootstrap estimation process, but resampling fewer than the full number of clusters C. I generate 1,000 cluster bootstrap samples each, sampling $\{2^{-8}C, 2^{-7}C, ...2^{-1}C, C\}$ clusters, estimate the IV models, and store the coefficients on the endogenous variables. Because of the slow speed of estimating causal forests, I only perform causal forest estimation for sample sizes up to $2^{-3}C$.[16]

For each cluster bootstrap sample I calculate mean absolute deviation from the full-sample parameters generated with the same method. Figures 11 and 12 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.

In both Figures 11 and 12, both the 2SLS and GroupSearch-based estimators have similar

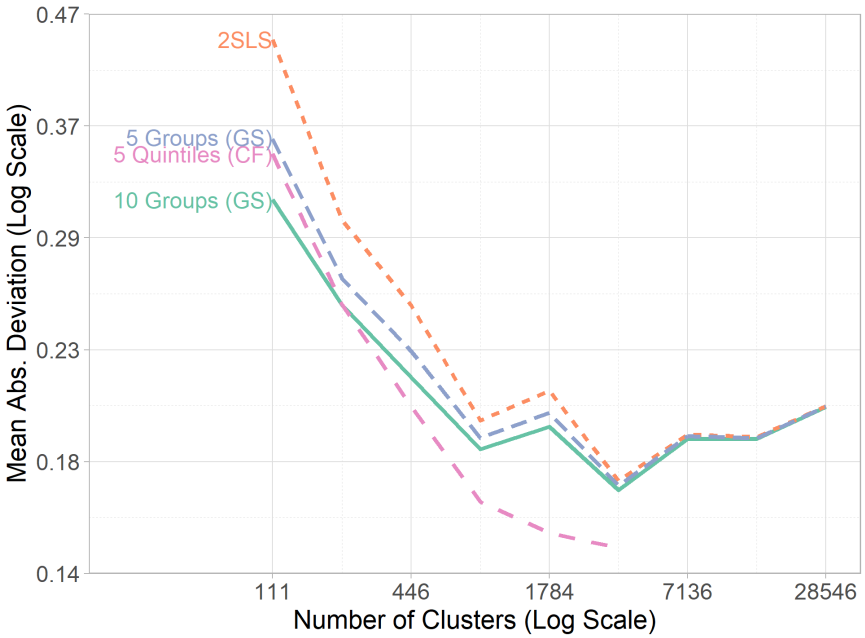
[16]Causal forest is feasible at larger sample sizes, for example in Table 1, but not at larger sample sizes 1,000 times.

Table 1: Replication of ABV Table 6 Without and With the Proposed 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
	Proposed 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
	Proposed 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 Angrist, Battistin, and Vuri (2017) 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 V. *p<0.1; **p<0.05; ***p<0.01

Figure 11: Performance in Replication of ABV Monitored × Class Size



Deviation is relative to the full-sample estimate in Table 1. At each sample size, 1,000 cluster-bootstrap samples are drawn. GroupSearch (GS) estimates use five (ten) groups based on the best of 100 randomly selected groupings. Causal Forest (CF) estimates use default settings in the R grf package, and quintiles of estimated effects are used. Causal Forest is only estimated for smaller samples due to computational limitations.

Figure 12: Performance in Replication of ABV Not Monitored × Class Size

Deviation is relative to the full-sample estimate in Table 1. At each sample size, 1,000 cluster-bootstrap samples are drawn. GroupSearch (GS) estimates use five (ten) groups based on the best of 100 randomly selected groupings. Causal Forest (CF) estimates use default settings in the R grf package, and quintiles of estimated effects are used. Causal Forest is only estimated for smaller samples due to computational limitations.

performance in the largest samples. However, all SLATE estimators tested outperform 2SLS in smaller samples. The version using causal forest to generate first-stage groups performs particularly well, and 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 proposed estimators to improve performance considerably, even in this setting where samples are relatively large and, as can be seen in Table 1, the instrument is very strong and so 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 proposed 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.

VI. 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 Miguel and Satyanath (2011) to Sarsons (2015) 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 show that studies often suffer from inadequate power (Young, 2018), and heightened sensitivity to heteroskedasticity and clustering (Andrews et al., 2019).

The reconstruction of IV necessarily must proceed on both fronts. Theoretical improve-

ments can come from stricter evaluation of exclusion restrictions as well as a series of new IV estimators that, at least in some contexts, weaken the reliance on validity (Kolesár et al., 2015; Windmeijer et al., 2018). Versions of the IV estimator that make statistical improvements under small samples or weak instruments already exist, especially under homoskedasticity, but are not applied at anywhere near a universal scale, even in top publications (see Andrews et al. 2019 for a review, as well as Chao and Swanson 2005 for the related literature on estimation with many weak instruments). Statistical improvements can come from more consistent application of methods robust to weak instruments.

This paper examines the implications of heterogeneity in the impact of an excluded instrument on an endogenous variable in instrumental variables estimation. I then introduce an estimation approach that incorporates heterogeneity in the first-stage estimate, reducing small-sample bias when the underlying effect is heterogeneous. This approach identifies 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).

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 estimation 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 paper 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 in poor conditions. 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 selection, although results will improve further

using causal forest or another method that uses covariates to model effect heterogeneity precisely. Further, the group SLATE estimator provides improved performance relative to 2SLS under heteroskedasticity, even though it is not derived with heteroskedasticity in mind, while many small-sample robust estimators rely on homoskedasticity (Andrews et al., 2019).

SLATE is also capable of improving robustness to monotonicity violations, at least under some conditions. Standard IV estimation, and its small-sample-robust variantions, are not robust to violations of monotonicity, and rely on assuming that monotonicity holds.

In addition to these general benefits of using a group-interaction SLATE estimator, right now is an opportune time to introduce the modeling of first-stage heterogeneity. The proposed 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. Top-k τ-Path does not perform particularly well in my simulation, but causal forest considerably approves performance in an applied context. These advances, which are still developing, make the SLATE estimators more powerful.

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 exclusion restrictions are looked upon with increasing skepticism. Still, IV is still used in cases where exclusion restrictions may be considered more defensible, like in fuzzy regression discontinuity, measurement error, 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.

VII. REFERENCES

Andrews, I., Stock, J. H., Sun, L., 2019. Weak Instruments in Instrumental Variables Regression: Theory and Practice. Annual Review of Economics 11 (1), 727–753.

Angrist, J. D., Battistin, E., Vuri, D., 2017. In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno. American Economic Journal: Applied Economics 9 (4), 216–249.

Angrist, J. D., Imbens, G. W., 1995. Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity. Journal of the American Statistical Association 90 (430), 431–442.

Angrist, J. D., Imbens, G. W., Rubin, D. B., 1996. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association 91 (434), 444–455.

Angrist, J. D., Lavy, V., 1999. Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. The Quarterly Journal of Economics 114 (2), 533–575.

Athey, S., Imbens, G., 2016. Recursive Partitioning for Heterogeneous Causal Effects. Proceedings of the National Academy of Sciences 113 (27), 7353–7360.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized Random Forests. The Annals of Statistics 47 (2), 1148–1178.

Bamattre, S., Hu, R., Verducci, J. S., 2017. Nonparametric Testing for Heterogeneous Correlation. In: Ahmed, S. E. (Ed.), Big and Complex Data Analysis: Methodologies and Applications. Contributions to Statistics. Springer International Publishing, Cham, pp. 229–246.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. Journal of Economic Perspectives 28 (2), 29–50.

Caloiaro, A., 2019. Topk tau-path. https://github.com/acaloiaro/topk-taupath, accessed: 2019-09-02.

Chao, J. C., Swanson, N. R., 2005. Consistent Estimation with a Large Number of Weak Instruments. Econometrica 73 (5), 1673–1692.

Fuller, W. A., 1977. Some Properties of a Modification of the Limited Information Estimator. Econometrica 45 (4), 939–953.

Ginestet, C. E., 2016. SteinIV: Semi-Parametric Stein-Like Estimator with Instrumental Variables. https://CRAN.R-project.org/package=SteinIV, accessed: 2019-09-27.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., Swanson, N. R., 2012. Instrumental Variable Estimation with Heteroskedasticity and Many Instruments. Quantitative Economics 3 (2), 211–255.

Heckman, J. J., Urzua, S., Vytlacil, E., 2006. Understanding Instrumental Variables in Models with Essential Heterogeneity. The Review of Economics and Statistics 88 (3), 389–432.

Heckman, J. J., Vytlacil, E. J., 2007. Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. Handbook of Econometrics 6 (Part B), 4779–4874.

Imbens, G. W., Angrist, J. D., 1994. Identification and Estimation of Local Average Treatment Effects. Econometrica 62 (2), 467–475.

Imbens, G. W., Wooldridge, J. M., 2009. Recent Developments in the Econometrics of Program Evaluation. Journal of Economic Literature 47 (1), 5–86.

Jiang, Y., Kang, H., Small, D., Zhao, Q., 2017. ivmodel: Statistical Inference and Sensitivity Analysis for Instrumental Variables Model. https://CRAN.R-project.org/package=ivmodel, accessed: 2019-09-27.

Kasy, M., 2014. Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment. The Review of Economic Studies 81 (4), 1614–1636.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., Imbens, G. W., 2015. Identification and Inference With Many Invalid Instruments. Journal of Business & Economic Statistics 33 (4), 474–484.

Miguel, E., Satyanath, S., 2011. Re-examining Economic Shocks and Civil Conflict. American Economic Journal: Applied Economics 3 (4), 228–232.

Nelson, C. R., Startz, R., 1990. The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument is a Poor One. The Journal of Business 63 (1), S125–S140.

Raudenbush, S. W., Bryk, A. S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. SAGE.

Sampath, S., Caloiaro, A., Johnson, W., Verducci, J. S., Sep. 2015. The Top-K Tau-Path Screen for Monotone Association. arXiv:1509.00549 [stat].

Sampath, S., Caloiaro, A., Johnson, W., Verducci, J. S., 2016. The Top-K Tau-Path Screen for Monotone Association in Subpopulations. Wiley Interdisciplinary Reviews: Computational Statistics 8 (5), 206–218.

Sampath, S., Verducci, J. S., 2013. Detecting the End of Agreement between Two Long Ranked Lists. Statistical Analysis and Data Mining: The ASA Data Science Journal 6 (6), 458–471.

Sarsons, H., 2015. Rainfall and Conflict: A Cautionary Tale. Journal of Development Economics 115, 62–72.

Staiger, D., Stock, J. H., May 1997. Instrumental Variables Regression with Weak Instruments. Econometrica; Evanston 65 (3), 557–586.

Wager, S., 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association 113 (523).

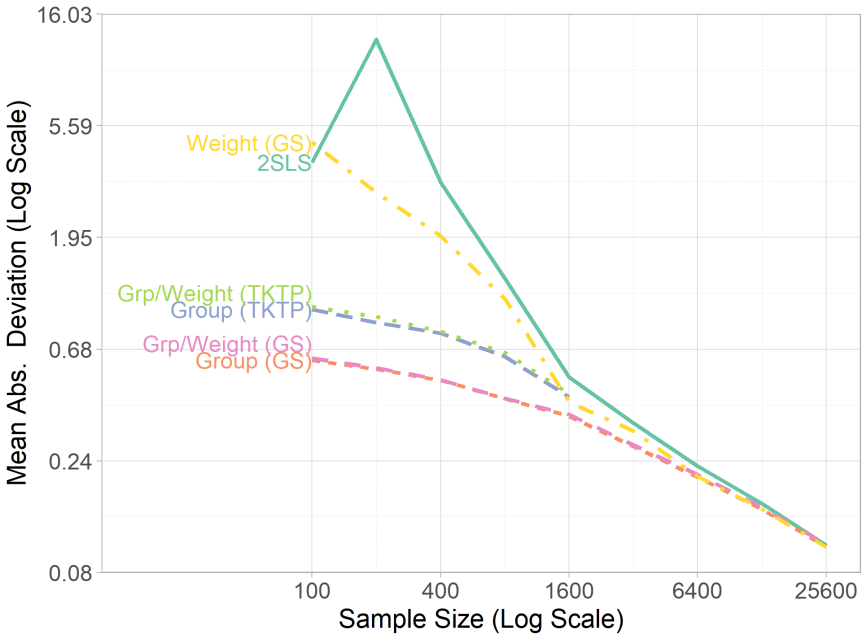
Windmeijer, F., Farbmacher, H., Davies, N., Smith, G. D., Jul. 2018. On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments. Journal of the American Statistical Association 114 (527), 1339–1350.

Wooldridge, J. M., 2010. Econometric Analysis of Cross Section and Panel Data, 2nd Edition. MIT Press, Cambridge, MA.

Young, A., 2018. Consistency without Inference: Instrumental Variables in Practical Application, unpublished.

Appendix A. Appendix

Figure A.13: Performance Using Feasible Estimation - Linear Specification



At each sample size, 1,000 random samples are drawn. GroupSearch (GS) estimates use a grouping of four groups from the best of 100 random groupings generated. Top-K τ-Path (TKTP) uses TKTP to identify groups in which z and x have positive, negative, or null relationships, respectively. TKTP is only run for smaller samples due to computational limitations. For Weight variants, first stage coefficients are estimated using groups, and then those coefficients are used to generate weights. See Section IV.i for data-generating process.