Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Instrumental variable estimation with first-stage heterogeneity[☆]

Alberto Abadie [a], Jiaying Gu [b], Shu Shen [c,*]

[a] Department of Economics, MIT, United States of America
[b] Department of Economics, University of Toronto, Canada
[c] Department of Economics, University of California, Davis, United States of America

ARTICLE INFO

Article history:
Received 13 November 2019
Received in revised form 11 January 2023
Accepted 2 February 2023
Available online xxxx

MSC:
C26

Keywords:
2SLS
First-stage heterogeneity
Asymptotic MSE
Pre-testing

ABSTRACT

We propose a simple data-driven procedure that exploits heterogeneity in the first-stage correlation between an instrument and an endogenous variable to improve the asymptotic mean squared error (MSE) of instrumental variable estimators. We show that the resulting gains in asymptotic MSE can be quite large in settings where there is substantial heterogeneity in the first-stage parameters. We also show that a naive procedure used in some applied work, which consists of selecting the composition of the sample based on the value of the first-stage t-statistic, may cause substantial over-rejection of a null hypothesis on a second-stage parameter. We apply the methods to study (1) the return to schooling using the minimum school leaving age as the exogenous instrument and (2) the effect of local economic conditions on voter turnout using energy supply shocks as the source of identification.

1. Introduction

While most of the methodological literature on instrumental variable (IV) methods assumes homogeneity in the first-stage parameters, empirical applications of IV estimators often involve settings where the strength of the instruments varies depending on the composition of the sample. Take the example of instruments based on policy changes as natural experiments. In this type of setting, variations in details of the policy or its level of enforcement across regions/jurisdictions may induce first-stage heterogeneity (e.g., Oreopoulos, 2006). Even if the policy intervention does not vary across regions, the identification strength of the instrumental variable may vary with the characteristics of the regions or because of differences in compliance rates across regions (e.g., Jackson et al., 2016) or demographic groups (e.g., Lleras-Muney, 2005, Stephens and Yang, 2014, Currie and Moretti, 2003).

In this article, we show that ignoring first-stage heterogeneity in IV models results in inefficient estimators, and propose IV estimators that improve precision over existing methods by addressing potential heterogeneity in the strength of the instruments.

In empirical studies in economics, it is common to select the sample on the basis of the strength of the instrument. For example, in the literature on the return to compulsory schooling, researchers often focus on Whites and/or early cohorts because data suggest that Blacks and more recent cohorts are weakly affected by changes in compulsory schooling

[☆] Whitney Newey's research, mentoring, and intellectual leadership has profoundly influenced our careers. For that, and for the many and always fascinating conversations on econometrics, academics, and life(!), we are immensely grateful to him. Thank you, Whitney!.

[*] Corresponding author.
E-mail addresses: abadie@mit.edu (A. Abadie), jiaying.gu@utoronto.ca (J. Gu), shushen@ucdavis.edu (S. Shen).

https://doi.org/10.1016/j.jeconom.2023.02.005
0304-4076/© 2023 Elsevier B.V. All rights reserved.

laws (see Lleras-Muney, 2005; Stephens and Yang, 2014).[1] Currie and Moretti (2003) uses county-level variation in college availability to study of the effect of mother's education on birth outcomes, but excludes Black mothers from the sample. The authors explain that, in their data, Black women are not as strongly affected in their educational level as White women by college availability. In a fuzzy RD study on the effect of publicizing workplace safety and health violations on outcomes of neighboring facilities, Johnson (2020) excludes two regions from the sample because data suggest low adherence to the RD cut-off rule in these two regions, resulting in a weak first-stage. Similarly, Cervellati et al. (2014) argues that the instrument used in an influential article by Acemoglu et al. (2008) on the effect of national income on democracy is weak for a sample of non-colonies, and focus their analysis on the sample of former colonies.

The first contribution of this article is to show that sample selection based on the first-stage correlation between an instrument and an endogenous variable using a fixed selection cut-off produces invalid inference for the two-stage least squares (2SLS) estimators. It produces second-stage IV estimators that can be severely biased, and second-stage t-statistics that can be too large under the null hypothesis in significance tests. Using different samples for sample selection in the first stage and 2SLS estimation in the second stage (e.g., the U.S. analysis in Altmejd et al., 2021) ameliorates these issues but results in inefficient estimators, as we discuss below.

An alternative approach employs variation in the strength of first-stage identification across groups of observations by interacting excluded instruments with group indicators. For example, in a study of the effect of air pollution on health outcomes, Deryugina et al. (2019) interact wind direction with pollution-monitor geo-cluster indicators to instrument for air pollution. Jackson et al. (2016) use a natural experiment of school finance reforms in the U.S. to investigate the effect of school spending on student outcomes. In one of their specifications they interact cohort and district group indicators to capture variation in the identification strength of the reform. Dix-Carneiro and Kovak (2017) interact excluded instruments with year dummies in a study of the effect of trade liberalization on Brazilian local labor markets. Allowing for time-varying first-stage coefficients, Pascali (2017) uses the introduction of steamships to identify the causal effect of globalization on economic development. This type of estimation strategy, which we call the fully-interacted method, is first-order efficient for models with groupwise first-stage heterogeneity under proper assumptions. Yet, inference under this procedure may be misleading because of many-IV bias, especially when the number of groups is large. In the above-mentioned studies, the total number of interacted instruments ranges from around twenty to over one hundred.

In this article, we propose a simple data-driven procedure that exploits heterogeneity in the first-stage correlation between an instrument and an endogenous variable to improve the asymptotic mean squared error (MSE) of 2SLS estimators. We consider a setting where the strength of an instrument varies across groups of the population defined by observables. If first-stage instrument strength is known for each population group, weighted 2SLS with weights reflecting the strength of the instrument in each group would be optimal under the assumption of homoskedasticity. In practice, IV strength is not known. Under our model set-up, weighted 2SLS with estimated weights is equivalent to 2SLS interacting the instrument with the full set of group dummy variables. Our proposed estimator uses tests of first-stage instrument relevance at the group level to select the sample, improving upon the fully interacted estimator. In our procedure, the cut-off value for first-stage testing is adaptively chosen to minimize the asymptotic MSE of the second-stage estimator. We use sample splitting following, for example, Chernozhukov et al. (2017, 2018) and Wager and Athey (2018) in our adaptive procedure to separate first-stage testing from second-stage estimation, reducing the asymptotic bias of the estimator.

Our set-up assumes a homogeneous second-stage to facilitate the comparison of different estimation approaches under the asymptotic MSE framework. However, when the second stage is heterogeneous, our proposed estimator has an interpretation as a weighted average causal effect. The weighting formula suggests important advantages of interacting both the external instrument and exogenous regressors with a full set of group indicators and reveals that groups with a weak first stage contribute dis-proportionally little to the final estimator. The analysis of heterogeneity provides additional motivation (beyond asymptotic MSE optimality) for the data-driven sample selection approach proposed in this article. Many competing efficient estimators with the same order in the higher order term of the MSE formula, including the limited information likelihood estimator (LIML) and the bias-corrected 2SLS estimator, do not have a weighted average causal effect interpretation under treatment effect heterogeneity. For the fully-interacted specification, the unbiased jackknife instrumental variables estimator (UJIVE) estimator proposed in Kolesár (2013) has the same weighted average causal effect interpretation as our baseline full-sample fully-interacted 2SLS estimator. We pursue a data-driven sample-selection extension of the fully-interacted 2SLS rather than UJIVE because our targeted empirical application on the return to compulsory schooling has a very large sample size and 2SLS is less computationally expensive than the jackknife instrumental variables estimator (JIVE, Phillips and Hale, 1997; Angrist et al., 1999) and its variants.

Our proposed methodology builds on the pioneering work of Donald and Newey (2001) on higher-order MSE expansion for IV estimators. For a wide range of cut-off values, our proposed estimators have the same first-order asymptotic distribution. We analyze the higher-order MSE behavior of the estimators, and propose a data-driven selector of cut-off values designed to minimize higher-order MSE.

[1] Footnote 44 of Lleras-Muney (2005) explains the exclusion of Blacks: "Lleras-Muney (2002) shows, for example, that the laws affected whites but not blacks". Stephens and Yang (2014) explains the exclusion of Blacks and the more recent cohorts: "the evidence on the efficacy of compulsory schooling laws is far more substantial for these cohorts than for more recent birth cohorts. Our analysis focuses on whites since we find no evidence supporting the efficacy of compulsory schooling laws for blacks in our sample".

The key difference between the MSE expansion literature (e.g., Donald and Newey, 2001; Okui, 2009; Kuersteiner and Okui, 2010) and the instrument selection strategy in the machine learning literature (e.g., Belloni et al., 2012; Chernozhukov et al., 2018) is that the IV selection criteria of the former is based on the asymptotic MSE of the second-stage estimator, while that of the latter is based on first-stage fitting. When there is only a vanishing proportion of groups with weak first-stage signals, our proposed estimator is asymptotically equivalent to the split-sample IV lasso method with groupwise transformed instruments, because both methods will select all groups with non-zero first-stage signals asymptotically. Our paper is also related to independent work by Coussens and Spiess (2021), who consider a setting with a randomized binary instrument, and propose to reweigh observations based on first-stage fit. Both our method and the method proposed in Coussens and Spiess (2021) utilize first-stage heterogeneity to improve the precision of second-stage IV estimation. Aside from that, the model, scope, and estimation procedure in Coussens and Spiess (2021) differ substantially from ours. On the one hand, our model considers first-stage heterogeneity at the group level, while Coussens and Spiess (2021) considers heterogeneity in first-stage coefficients across individual observations (with the strength of the first-stage explained by observable characteristics of the observations). On the other hand, while Coussens and Spiess (2021) concentrates on a setting with a randomized binary instrument, we allow for non-binary IV and the presence of exogenous controls. In addition, while Coussens and Spiess (2021) assume that the first-stage is strong for all observations, we allow for the presence of groups with zero or weak first-stage, and we select the groups that make up the sample to minimize second-stage MSE.

On the empirical side, this article contributes results to the return to schooling literature and to research that utilizes energy supply shocks to instrument for local economic conditions. Our first empirical application reanalyzes the data in Stephens and Yang (2014), who argue in favor of controlling for regional cohort fixed effects in studies of the return to schooling. Once they control for regional cohort fixed effects, Stephens and Yang (2014) obtain IV estimates of the return to schooling that are not statistically significant.

We find that, after taking into account first-stage heterogeneity across geographic regions and demographic groups, our proposed procedure consistently produces statistically significant estimates of 3–4 percent for the effect of an additional year of schooling on wages for the specification with regional cohort fixed effects. These results are estimated for adaptively selected groups of White males and White females mostly in the Northeast, Midwest, and South of the U.S.

Our second empirical application revisits the Charles and Stephens (2013) study of the effect of local labor market variables on voter turnout in U.S. elections, with labor market variables instrumented by employment shocks in the oil and coal industries. The main IV specification in Charles and Stephens (2013) uses the 1974 County Business Patterns data (CBP) to measure county-level employment in the oil and coal industries. Although the 1974 dataset contains detailed industry-level information at the county level, instruments based on these data may not provide completely exogenous variation for the 1969–2000 estimation window in Charles and Stephens (2013). As a robustness check, Charles and Stephens (2013) use the 1967 CBP data. The 1967 CBP specification is based on a cleaner exclusion restriction, but produces a weaker first-stage than the specification based on the 1974 data because the 1967 CBP measures county-level data for the entire mining industry. We find that full-sample 2SLS and 2SLS restricted to states with substantial shares for the oil and coal industries produce statistically insignificant coefficients with the 1967 CBP instrument. However, more efficient estimators, including our proposed adaptive procedure, produce negative and statistically significant effects of local market activity on voter turnout. Applying our procedure to the 1967 CBP data generates results that are qualitatively similar to those reported in Charles and Stephens (2013) for the 1974 CBP data.

The remainder of this article is organized as follows. Section 2 sets up a simultaneous equation model where the correlation between the instrument and the endogenous variable could be strong, weak, or zero for different population subgroups. We discuss the asymptotic properties of the existing methods and the drawbacks of the naive direct selection approach often used in applied work. In Section 3, we study the behavior of a modified selective IV estimator that is consistent and efficient under mild conditions. We analyze the asymptotic MSE of the proposed estimator as a function of a first-stage selection cut-off, and propose a data-driven procedure to estimate the cut-off and construct a data-driven adaptive IV estimator. In Section 4, we use simulations to confirm the MSE improvement of our proposed adaptive estimator. In Section 5, we report the results from empirical applications to the compulsory schooling data of Oreopoulos (2006) and the voter turnout data of Charles and Stephens (2013). Section 6 concludes.

2. Model set-up and existing methods

2.1. Model set-up

As we discuss in the introduction, it is often the case in applied settings that the correlation between an endogenous variable and an instrument is heterogeneous across different population groups. Consider a simultaneous equation model with a heterogeneous first stage, where the instrument is strong for some population groups, weak for some other groups, and uncorrelated with the endogenous variable for the rest. This model is a natural specification for a variety of economic applications. For example, in literature on the return to compulsory schooling, economists compile information from multiple natural experiments (e.g., state laws that shift minimum school dropping age) to create an instrument (e.g., the minimum school dropping age an individual faced at the age of 14). This instrument is used to estimate the effect of an endogenous variable (years of education) on the outcome (wages). Effective policies—that is, policies that affect the years of education—make the instrument correlated with the endogenous variable, while ineffective policies undermine this correlation.

We posit a simultaneous equation model with one endogenous regressor, W, and one instrument, $\tilde{Z}$. Suppose that we observe N individuals, who are divided into G groups. We know which group each individual belongs to. We assume that for each individual i in group g, we have

$$\begin{aligned} Y_{ig} &= \beta W_{ig} + X_{ig}\theta_g + u_{ig}, \\ W_{ig} &= \rho_g \tilde{Z}_{ig} + X_{ig}\gamma_g + v_{ig}, \end{aligned} \tag{1}$$

where X_{ig} is a vector of covariates of dimension $1 \times d$. Within each group, $(\tilde{Z}_{ig}, X_{ig}, u_{ig}, v_{ig})$ are i.i.d. and there is potentially a non-zero correlation between u_{ig} and v_{ig}. The model has heterogeneous first-stage coefficients across groups as well as group-specific effects of exogenous regressors. In empirical research, groups could be determined by observables like geographic regions, ethnic groups, etc. To facilitate the comparison among different estimators in an asymptotic MSE framework, we assume that β is a constant for our benchmark model. In Section 2.2.2, we discuss the interpretation of existing and proposed estimators under heterogeneity in causal effects.

After residualizing the exogenous variables from the instrument (groupwise) and writing the model in matrix form, we have

$$\begin{aligned} Y_g &= \beta W_g + X_g\theta_g + u_g, \\ W_g &= Z_g\rho_g + X_g\omega_g + v_g, \end{aligned}$$

where Y_g, W_g, $\tilde{Z}_g$, u_g, v_g are vectors of length n_g, X_g is a matrix of dimension $n_g \times d$, $Z_g = M_{X_g}\tilde{Z}_g$ where $M_{X_g} = I - X_g(X_g'X_g)^{-1}X_g'$, and $\omega_g = \gamma + (X_g'X_g)^{-1}(X_g'\tilde{Z}_g)\rho_g$. By construction, $Z_g'X_g = 0$. The following assumption provides regularity conditions.

Assumption 1.

1. Data Design: Observations are independent across groups and i.i.d. conditional on grouping. There exist positive and finite $\underline{c}$ and $\bar{c}$ such that $\underline{c}N/G \leq n_g \leq \bar{c}N/G$ for all $g = 1, 2, \ldots, G$ and $G/N \to 0$ and $N \to \infty$.
2. One-sided First-stage Relationship: There exist constants $a_1, \ldots, a_G$, and positive and finite $\underline{\rho}$ and $\bar{\rho}$ such that $\underline{\rho} \leq a_g < \bar{\rho}$ for all $g = 1, \ldots, G$. Groups with irrelevant IV are defined as $\mathcal{G}_0 = \{g : \rho_g = 0\}$, groups with strong IV are defined as $\mathcal{G}_{+,s} = \{g : \rho_g = a_g\}$, and groups with weak IV are defined as $\mathcal{G}_{+,w} = \{g : \rho_g = a_g/\sqrt{n_g}\}$. We further denote $G_0 = |\mathcal{G}_0|$, $G_{+,s} = |\mathcal{G}_{+,s}|$, $G_{+,w} = |\mathcal{G}_{+,w}|$, and let $\mathcal{G}_+ = \mathcal{G}_{+,w} \cup \mathcal{G}_{+,s}$ and $G_+ = G_{+,w} + G_{+,s}$.
3. Finite Moments: Let $k_g = E[\eta_g'\eta_g/n_g]$ where η_g is the error vector after projecting $\tilde{Z}_g$ linearly onto X_g. There exist positive and finite $\underline{k}$ and $\bar{k}$ such that $\underline{k} \leq k_g \leq \bar{k}$ for all $g = 1, \ldots, G$. In addition, there exists a positive and finite constant that bounds $E[\tilde{Z}_{ig}^8]$ and $E[X_{ig}^8]$ uniformly across all $g = 1, \ldots, G$.
4. Error Terms: For all $g = 1, \ldots, G$, $(u_{ig}, v_{ig})|(\tilde{Z}_{ig}, X_{ig})$ have a common distribution with mean 0 and non-singular variance–covariance matrix $\begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}$. In addition, there exists a positive and finite constant that bounds $E[v_{ig}^8]$ uniformly across all $g = 1, \ldots, G$.
5. Non-trivial Presence of Strong Groups: When G is fixed, $G_{+,s} > 0$. When $G \to \infty$, $G_{+,s}/G \to b > 0$.

Assumption 1.1 allows for unbalanced sample sizes by group, but requires that all groups have sample sizes of the same order, both in the fixed and the growing G cases.

Assumption 1.2 requires that the instrument affects the endogenous regressor in the same direction across all groups. It is adopted here for notational simplicity but is also in line with the monotonicity condition in the local average treatment effect (LATE) literature (see Angrist and Pischke, 2009 for a review). Without loss of generality, we assume that first-stage effects are non-negative similar to Andrews and Armstrong (2017). Assumption 1.2 is a natural assumption in our two empirical applications, where compulsory schooling laws instrument for years of education and energy supply shocks instrument for local economic conditions. Assumption 1.2 allows for weak first-stage relationships. This is more general than the well-separated case (where first-stage coefficients are either zero or they are bounded away from zero). As is common in the weak IV literature, we say that a first-stage relationship is weak if the first-stage coefficient is of order one over the square root of the (within group) sample size.

Assumption 1.3 requires the instrument to have non-trivial variation in each group. In practice, groups with zero or very small variation in the instrument can be dropped in advance, if necessary. Assumption 1.3 allows the variance of the instrument to be heterogeneous across groups, which could be an important feature of some empirical applications, including the two examples we study later in the article.

Assumption 1.4 imposes exclusion restrictions and the homoskedasticity condition on the distribution of error terms. These assumptions are commonly adopted in the literature. Assumption 1.5 is required for the identification of β. Similar assumptions of strong identification are often employed in the IV literature. For example, Okui (2009) and Cheng et al. (2019) assume that researchers have prior knowledge about a subset of informative or strong instruments. In this article, we require non-trivial presence of population subgroups with strong instruments, but we do not require prior knowledge of the identity of the relevant subgroups.

2.2. Existing methods

Let ℓ_g denote a vector of n_g ones for $g = 1, 2, \dots, G$ and ℓ denote a vector of N ones. Let Y, W, X, $\tilde{Z}$, Z, u, v, and ℓ be vectors or matrices with row size N that stack all group vectors Y_g, W_g, X_g, $\tilde{Z}_g$, Z_g, u_g, v_g, and ℓ_g, respectively. For any full-rank matrix A, let $P_A = A(A'A)^{-1}A'$ and $M_A = I - P_A$. In this section, we discuss IV estimators that are often used in empirical studies with potential first-stage heterogeneity.

2.2.1. Pooled and fully interacted 2SLS

Let $\tilde{D}$ be the $N \times G$ block diagonal matrix of $\tilde{Z}_1, \tilde{Z}_2, \dots, \tilde{Z}_G$, D the $N \times G$ block diagonal matrix of $Z_1, \dots, Z_G$, D_X the $N \times G$ block diagonal matrix of $X_1, \dots, X_G$, and D_ℓ the $N \times G$ block diagonal matrix of $\ell_1, \dots, \ell_G$. D_ℓ is the set of group indicators and $\tilde{D}$ (or D, D_X) includes all interaction terms between $\tilde{Z}$ (or Z, X) and the set of group indicators. The most commonly used IV estimators in empirical studies with potential first-stage heterogeneity across groups are: *(i)* the pooled 2SLS estimator,

$$\hat{\beta}_{pool} = (Z'W)^{-1}Z'Y,$$

which ignores group membership, and *(ii)* the fully-interacted 2SLS estimator,

$$\hat{\beta}_{int} = \left(W'P_D W\right)^{-1} W'P_D Y,$$

that accounts for the groupwise heterogeneity in model (1) by interacting the instrument with a full set of group membership indicators. The fully-interacted estimator could also be written as $\hat{\beta}_{int} = \left(\sum_{g=1}^{G} \hat{\rho}_g Z_g' W_g\right)^{-1} \sum_{g=1}^{G} \hat{\rho}_g Z_g' Y_g$, where $\hat{\rho}_g = (Z_g'Z_g)^{-1}(Z_g'W_g)$ is the groupwise first-stage estimator for ρ_g. Using the groupwise transformed instrument Z is equivalent to allowing for groupwise slopes for the exogenous regressor X, which (as we discuss in Section 2.2.2) can be important for the interpretability of the estimator even if the true slopes of X are homogeneous across groups.

Let $p_g = n_g/N$ for all $g = 1, 2, \dots, G$. Under Assumption 1, the pooled estimator $\hat{\beta}_{pool}$ satisfies

$$\sqrt{N}\left(\hat{\beta}_{pool} - \beta\right)/s_p \Rightarrow N(0, 1), \quad s_p = \sigma_u \Big/ \sqrt{\left(\sum_{g=1}^{G} \rho_g k_g p_g\right)^2 \Big/ \left(\sum_{g=1}^{G} k_g p_g\right)}.$$

Under Assumption 1 and the additional rate condition $G^2/N \to 0$, the fully-interacted estimator $\hat{\beta}_{int}$ satisfies

$$\sqrt{N}\left(\hat{\beta}_{int} - \beta\right)/s_{int} \Rightarrow N\left(0, 1\right), \quad s_{int} = \sigma_u \Big/ \sqrt{\sum_{g=1}^{G} \rho_g^2 k_g p_g}.$$

Both estimators are consistent. The fully interacted estimator is more efficient since $s_{int} \le s_p$ by the Cauchy–Schwarz inequality. The equality holds if and only if the groupwise first-stage slope ρ_g is constant across groups.

The growth condition $G^2/N \to 0$ is required to guarantee that the asymptotic bias of $\hat{\beta}_{int}$ vanishes in the limit. The fully-interacted estimator, $\hat{\beta}_{int}$, has the same asymptotic distribution as the infeasible oracle 2SLS estimator using $Z_{inf} = (\rho_1 Z_1', \dots, \rho_G Z_G')'$ as the instrument and is hence efficient under homoskedasticity. If homoskedasticity is violated, efficient estimation of β would require a GLS-type of reweighing involving estimated variance of the second-stage error term. In practice, however, the fully-interacted estimator may suffer from many-IV bias as is discussed in Bekker (1994), Bound et al. (1995), Staiger and Stock (1997), and Stock and Yogo (2005), among many others.

2.2.2. Interpretation under second-stage effect heterogeneity

The popularity of 2SLS in empirical research is in part due to the fact that, if treatment effects are heterogeneous, 2SLS has a weighted average treatment effect interpretation under some conditions. When the endogenous regressor and instrument are both binary, and there are no other exogenous covariates in the model, 2SLS estimates the average treatment effect of compliers (Imbens and Angrist, 1994). This subsection studies the interpretation of pooled and fully-interacted 2SLS estimators when a groupwise heterogeneous second-stage causal parameter is added to our model in (1). To facilitate the discussion, we temporarily simplify the exogenous regressor X to contain only the intercept. In the next section we bring back the general case.

Replace β in model (1) with β_g and assume $|\beta_g| \le \bar{\beta} < \infty$ for all $g = 1, \dots, G$. It is then easy to show that given the regularity conditions in Assumption 1 and the corresponding rate conditions (i.e., $G/N \to 0$ for $\hat{\beta}_{pool}$ and $G^2/N \to 0$ for $\hat{\beta}_{int}$),

$$\hat{\beta}_{pool} = \sum_{g=1}^{G} \frac{\rho_g V_g p_g}{\sum_{g=1}^{G} \rho_g V_g p_g} \beta_g + o_p(1), \quad \hat{\beta}_{int} = \sum_{g=1}^{G} \frac{\rho_g^2 V_g p_g}{\sum_{g=1}^{G} \rho_g^2 V_g p_g} \beta_g + o_p(1), \tag{2}$$

where $V_g = V[\tilde{Z}_{ig}]$. For both estimators, groups with larger variance in the instrument receive higher weights in the probability limit. The results in (2) are related to those in Angrist and Imbens (1995) and Abadie (2003), which establish a causal interpretation for 2SLS estimators under parameter heterogeneity (see, e.g., Theorem 3 in Angrist and Imbens, 1995, and Proposition 5.1 in Abadie, 2003).

The groupwise first-stage slope enters the weighting formula linearly for the pooled estimator but in a squared form for the fully interacted estimator. Although at first glance, the squared form may not appear intuitive, it reflects an advantage of the fully interacted estimator: the fully interacted estimator is invariant to groupwise rescaling of the instrument. For example, if the instrument in group g is multiplied by a factor a, the variance of the instrument in group g increases by a^2, but the first-stage slope coefficient ρ_g is divided by a only. As a result, groupwise changes in the scale of the instrument (resulting, for example, from standardizing different units of measurement of the instrument across groups/jurisdictions) change the interpretation of the pooled estimator but not of the fully interacted 2SLS estimator.

The interpretation of 2SLS estimates as weighted averages of causal effects motivates the use of the groupwise transformed instrument, Z, even for the case when the slope coefficients θ_g and γ_g in model (1) are assumed to be homogeneous across groups. In the absence of groupwise transforming of the instrument, the interpretation of the pooled and fully-interacted estimators becomes complicated. Intuitively, imposing the same intercept across groups allows groups with no first-stage identification, or $\rho_g = 0$, to influence the 2SLS estimator through their influence on the value of the intercept. For example, for a model with no exogenous variables other than the group indicators, the pooled 2SLS estimator with a universal intercept is $\hat{\beta}_{pool2} = \left(\tilde{Z}' M_\ell W\right)^{-1} \tilde{Z}' M_\ell Y$. Let $d_g = E[\tilde{Z}_{ig}^2]$ and $b_g = E[\tilde{Z}_{ig}]$. In the appendix, we show $\hat{\beta}_{pool2} = (\sum_{g=1}^G \rho_g p_g (d_g - b_g \sum_{s=1}^G b_s p_s))^{-1} (\sum_{g=1}^G \beta_g p_g (\gamma (b_g - \sum_{s=1}^G b_s p_s) + \rho_g (d_g - b_g \sum_{s=1}^G b_s p_s))) + o_p(1)$, where γ is the true intercept in the model. The first term on the right-hand side of last equation is not a weighted average (the factors that multiply β_g sum to one, as shown in the appendix, but they could be negative or larger than one), and even groups with $\rho_g = 0$ influence the estimator. Also, because of the role of γ in the previous formula, the estimator $\hat{\beta}_{pool2}$ is not invariant to recentering of the endogenous regressor, W.

The result in (2) provides additional motivation for a data-driven selection of strong groups in the fully-interacted 2SLS model. Because the first-stage slope, ρ_g, enters the fully-interacted 2SLS formula in (2) in a squared form, the contribution of different groups to the weighted average of the coefficients β_g can be extremely uneven, complicating the interpretation of the fully-interacted 2SLS estimate. Section 2.2.3 discusses the inferential pitfalls of naive first-stage sample selection. The results in the current section and Section 2.2.3 motivate a select-and-interact 2SLS procedure, which we propose in Section 3.

2.2.3. *Naive first-stage selection*

Applied researchers often employ a direct sample selection approach to obtain a strong first stage. If the selected sample is based on economic intuition (e.g., Fredriksson et al., 2013; Card et al., 2014), the selective IV approach may be legitimate. However, when the selection is based on first-stage regression results, the sample selection process invalidates the exclusion restriction at a rate that endangers the validity of post-selection IV inference.

To show the breakdown of inference after sample selection based on the strength of the instrument, we first formally describe the data-driven selective IV approach, which consists of running an IV regression using only the groups selected by testing $H_{0,g} : \rho_g = 0$ against the alternative $H_{a,g} : \rho_g > 0$, $g = 1, \ldots, G$. Let t_g be the t-statistic for group g, α_{FS} be a pre-determined and fixed significance level, and $c_{g,\alpha_{FS}}$ be the $(1 - \alpha_{FS})$ quantile of Student-t distribution with $n_g - 1$ degrees of freedom. Let $i_{g,\alpha_{FS}} = 1(t_g > c_{g,\alpha_{FS}})$. Assuming that at least one group is selected, the resulting estimator is

$$\hat{\beta}_{selp} = \left(\sum_{g=1}^G i_{g,\alpha_{FS}} Z_g' W_g\right)^{-1} \sum_{g=1}^G i_{g,\alpha_{FS}} Z_g' Y_g. \tag{3}$$

We refer to the estimator in (3) as the select-and-pool estimator.

The next theorem shows that the exclusion restriction is violated for the select-and-pool estimator at a rate that invalidates the conventional inference.

Theorem 1. *Suppose Assumption 1 holds and $\sigma_{uv} \neq 0$. Let $0 \leq \alpha_{FS} < 1/2$, then*

$$E\left[1\left(\sum_{g=1}^G n_g i_{g,\alpha_{FS}} > 0\right) \left|\sum_{g=1}^G i_{g,\alpha_{FS}} Z_g' u_g\right| \Big/ \sum_{g=1}^G n_g i_{g,\alpha_{FS}}\right] \geq a/\sqrt{N/G} + o(1/\sqrt{N/G}),$$

for some positive constant a.

Proof of the theorem is provided in the appendix.[2] The theorem has multiple implications. First, it implies that the select-and-pool method violates the exclusion restriction for any finite sample. This is because selection is based on the value of the first-stage t-statistic in each group, and a group is more likely to be selected when there is a large positive

[2] The proof applies to settings more general than those covered in Theorem 1. In particular, it allows for negative values for the constants a_g (relaxing Assumption 1.2) and groupwise heteroskedasticity (relaxing Assumption 1.4).

Table 1
Rejection rates of existing estimators.

	$G_{+,s}/G = 0.1$						$G_{+,s}/G = 0.3$					
	$\rho_{uv} = 0.25$			$\rho_{uv} = 0.5$			$\rho_{uv} = 0.25$			$\rho_{uv} = 0.5$		
	$\hat{\beta}_{pool}$	$\hat{\beta}_{int}$	$\hat{\beta}_{selp}$	$\hat{\beta}_{pool}$	$\hat{\beta}_{int}$	$\hat{\beta}_{selp}$	$\hat{\beta}_{pool}$	$\hat{\beta}_{int}$	$\hat{\beta}_{selp}$	$\hat{\beta}_{pool}$	$\hat{\beta}_{int}$	$\hat{\beta}_{selp}$
$G = 10$												
$n = 250$	0.007	0.082	0.036	0.030	0.264	0.100	0.018	0.074	0.047	0.042	0.165	0.080
$n = 500$	0.010	0.063	0.039	0.028	0.184	0.072	0.020	0.063	0.046	0.034	0.105	0.069
$n = 1000$	0.013	0.068	0.061	0.036	0.144	0.094	0.036	0.052	0.055	0.043	0.075	0.072
$G = 40$												
$n = 250$	0.013	0.209	0.073	0.036	0.683	0.186	0.036	0.122	0.061	0.043	0.391	0.111
$n = 500$	0.028	0.169	0.077	0.052	0.521	0.184	0.056	0.122	0.065	0.050	0.258	0.097
$n = 1000$	0.026	0.105	0.061	0.035	0.301	0.149	0.046	0.066	0.044	0.044	0.125	0.072
$G = 100$												
$n = 250$	0.025	0.414	0.115	0.044	0.971	0.341	0.048	0.237	0.072	0.051	0.737	0.182
$n = 500$	0.031	0.302	0.111	0.038	0.841	0.263	0.043	0.168	0.072	0.040	0.500	0.118
$n = 1000$	0.041	0.210	0.102	0.045	0.633	0.259	0.057	0.110	0.055	0.060	0.303	0.106
$G = 200$												
$n = 250$	0.031	0.698	0.190	0.038	0.999	0.532	0.043	0.435	0.111	0.040	0.948	0.308
$n = 500$	0.041	0.551	0.166	0.045	0.984	0.435	0.057	0.287	0.072	0.060	0.775	0.184
$n = 1000$	0.038	0.360	0.144	0.039	0.899	0.418	0.043	0.157	0.079	0.041	0.527	0.156

Note: The table reports rejection rates of the Wald test based on different estimators for $H_0 : \beta = 0$ among 1000 simulations with 5 percent nominal level. The data generating process is $X_{ig}, \tilde{Z}_{ig} \sim i.i.d.\ N(0,1)$, $(u_{ig}, v_{ig}) \sim N((0\ 0), (1\ \rho_{uv};\ \rho_{uv}\ 1))$, $W_{ig} = \rho_g \tilde{Z}_{ig} + X_{ig} + v_{ig}$, $Y_{ig} = \beta W_{ig} + X_{ig} + u_{ig}$ for $i = 1, 2, \ldots, n$, where $\beta = 0$, $\rho_g = 0.2$ for $g = 1, \ldots, G_{+,s}$ and $\rho_g = 0$ for $g > G_{+,s}$.

correlation between the instrument and the first-stage error term. Because the first and second-stage error terms are correlated, the select-and-pool procedure induces a violation of the exclusion restriction.

Violation of the exclusion restriction for any finite sample, however, does not necessarily imply inconsistency of IV. Nor does it imply that classic inference methods become invalid. An important previous literature has studied local to zero violations of the exclusion restriction, particularly for a regime correlation of order $1/\sqrt{N}$ (e.g. Staiger and Stock (1997), Berkowitz et al. (2008), and Guggenberger (2012) among others). In this regime of local violation, classical inference starts to fail for IV estimators. For instance Berkowitz et al. (2008) shows that 2SLS has a limiting distribution that no longer centers at the true parameter value under such local violation. Our result in Theorem 1 implies that, for fixed G, select-and-pool violates the exclusion restriction at a rate no smaller than $1/\sqrt{N}$ as long as not all groups have strong first-stage coefficients. When G grows together with N, Theorem 1 implies that the exclusion restriction of the select-and-pool estimator is violated at a rate worse than $1/\sqrt{N}$. Under such circumstances, the type I error of conventional t-tests based on the select-and-pool estimator converges to one.

Table 1 illustrates the over-rejection problem of the select-and-pool estimator. The data generating process (DGP) used for the simulations is described in the footnote of the table. As predicted by Theorem 1, the test based on the select-and-pool estimator over-rejects more severely when the number of groups grows, and the size distortion is not alleviated with the increase of sample size. When $G = 10$, the rejection rates of the select-and-pool estimator range from 6.9 to 10 percent when $\rho_{uv} = 0.5$. When $G = 100$, the rejection rate can be as high as 34 percent. The over-rejection problem also gets worse with increased model endogeneity and higher proportion of zero groups.

Table 1 also reports the finite-sample performance of the pooled and the fully interacted estimators. The pooled estimator controls size well. But it is also highly inefficient, as shown in Table A1 in the Appendix, which reports standard deviations for the different estimators. The fully interacted estimator suffers from many-IV bias with the size distortion increasing with the number of groups, the degree of endogeneity, and the first-stage weakness of the instrument (proportion of groups with irrelevant IV). In the DGP of this simulation experiment, many-IV bias is a finite sample problem, and the size distortion of $\hat{\beta}_{int}$ improves as the sample size grows. Table A2 in the Appendix provides a closer look at the finite-sample biases of the different estimators using the same data generating processes (DGPs) as in Table 1.

2.2.4. JIVE and its variants

The jackknife instrumental variables estimator (JIVE) of Phillips and Hale (1997) and Angrist et al. (1999) uses leave-one-out in the first stage to mitigate the many-IV bias of 2SLS. Kolesár (2013) proposes an unbiased jackknife instrumental variables estimator (UJIVE) that further corrects the many-covariates bias of JIVE.[3] Kolesár (2013) shows that UJIVE has the same weighted average causal effect interpretation as the corresponding 2SLS estimator under treatment effect heterogeneity (where the weights could be negative or larger than one depending on model specification). The result of Kolesár (2013) is in line with our analysis in Section 2.2.2. With first-stage groupwise heterogeneity, a fully-interacted

[3] Ackerberg and Devereux (2009) propose an improved JIVE (IJIVE) estimator also aimed to correct the many-covariates bias of JIVE.

version of UJIVE (interacting both the instrument and the set of exogenous regressors with group indicators) has the same interpretation as the fully-interacted 2SLS estimator.

In the next section, we propose a split-sample select-and-interact 2SLS procedure, and derive a first-stage selection criterion that is optimal with respect to second-stage asymptotic MSE. As shown in Theorems 1 and 2 of Kolesár (2013), the UJIVE estimator has the same limiting weighting scheme interpretation as the corresponding two-step IV estimator (for the same specification). We work with 2SLS because of its computational advantage relative to JIVE and its variants. In the simulation section, we compare the small sample performance of the proposed adaptive procedure to the performance of the UJIVE and IJIVE procedures.

3. Adaptive estimation

We have shown in the previous section that the fully-interacted 2SLS estimator has a simple and intuitive interpretation as a weighted average causal effect when model (1) is extended to allow for groupwise heterogeneity in the second-stage parameters. However, the fully-interacted 2SLS estimator could be subject to substantial many-IV bias in finite samples when the number of groups is large. It is, therefore, natural to ask if it is possible to construct a new estimator that preserves the intuitive interpretation of the fully-interacted 2SLS estimator and mitigates its many-IV bias problem, without large increases in variance. We next propose an estimator that satisfies these requirements and is amenable to classic asymptotic inference. Our estimation procedure selects a set of groups with strong first-stage effects, and uses the selected groups only to calculate a fully-interacted 2SLS estimator. When the causal effects are heterogeneous across groups, our procedure provides transparent information on the identity of the groups that contribute to the parameter estimated by fully-interacted 2SLS.

3.1. Split-sample select-and-interact 2SLS

We define the select-and-interact estimator, $\hat{\beta}_{sel,int}(\delta)$, in the same way as the fully-interacted estimator, except that only groups that pass a first-stage significance test are used to estimate β,

$$\hat{\beta}_{sel,int}(\delta) = \left(\sum_{g=1}^{G} \hat{\rho}_g Z_g' W_g 1(\hat{\mu}_g > \delta)\right)^{-1} \sum_{g=1}^{G} \hat{\rho}_g Z_g' Y_g 1(\hat{\mu}_g > \delta), \tag{4}$$

where $1(\hat{\mu}_g > \delta)$ is the selection rule with $\hat{\mu}_g = \hat{\rho}_g (Z_g' Z_g)^{1/2} = (Z_g' Z_g)^{-1/2} Z_g' W_g$, for some δ. When $\delta = -\infty$, the estimator reduces to the fully-interacted 2SLS estimator $\hat{\beta}_{int}$. If the interactions between Z and group indicators are pre-normalized to have unit variances (as is usual for regularized regression methods such as lasso and ridge) the selection rule $1(\hat{\mu}_g > \delta)$ is solely based on the magnitude of the first-stage slope coefficient estimator, $\hat{\rho}_g$. In this section, we examine the statistical properties of select-and-interact estimators when δ is a fixed constant. In Section 3.3 we consider the adaptive choice of δ based on an expansion of the MSE of the second-stage estimator.

Note that the estimator in (4) runs the second-stage regression using only data from the groups selected in the first stage. The estimator is identical to a full-sample 2SLS estimator of the second-stage causal parameter if D_X is used as the exogenous regressor and columns in the matrix D corresponding to the selected groups are used as excluded instruments. The drawback of using the full-sample 2SLS regression is that, although data from unselected groups do not affect the second-stage estimator itself, they affect the standard error calculation through the estimation of σ_u. Therefore, we choose to define the select-and-interact estimator as in (4) and carry out 2SLS only using data from selected groups.

We next propose a split-sample version of the select-and-interact 2SLS estimator. We first randomly split the data into two subsamples of equal proportions. We use superscripts a and b to refer to the observations in the two sample splits. Let N^a be the sample size of split a and N^b be the sample size of split b. Let $Z_g^a = M_{X_g^a}\tilde{Z}_g^a$, $\hat{\rho}_g^a = \left((Z_g^a)' Z_g^a\right)^{-1} (Z_g^a)' W_g^a$, $\hat{\mu}_g^a = \left((Z_g^a)' Z_g^a\right)^{-1/2} (Z_g^a)' W_g^a$, and define similar terms for sample split b. Let

$$\hat{\beta}^a(\delta) = \left(\sum_g \hat{\rho}_g^b (Z_g^a)' W_g^a 1(\hat{\mu}_g^b \geq \delta)\right)^{-1} \sum_g \hat{\rho}_g^b (Z_g^a)' Y_g^a 1(\hat{\mu}_g^b \geq \delta),$$

$$\hat{\beta}^b(\delta) = \left(\sum_g \hat{\rho}_g^a (Z_g^b)' W_g^b 1(\hat{\mu}_g^a \geq \delta)\right)^{-1} \sum_g \hat{\rho}_g^a (Z_g^b)' Y_g^b 1(\hat{\mu}_g^a \geq \delta),$$

and

$$\hat{\beta}_{sssel,int}(\delta) = \left(\hat{\beta}^a(\delta) + \hat{\beta}^b(\delta)\right)/2. \tag{5}$$

$\hat{\beta}^a(\delta)$ and $\hat{\beta}^b(\delta)$ use one of the splits for first-stage instrument selection and reweighting and the other split for second-stage estimation. As we show below, by averaging across $\hat{\beta}^a(\delta)$ and $\hat{\beta}^b(\delta)$, the repeated split-sample select-and-interact estimator defined in (5) preserves efficiency. Averaging of split-sample estimator has been previously used by Belloni et al. (2012) and others.

The following Assumption gives a range condition for δ.

Assumption 2 (*Range of* δ)**.** The thresholding value $\delta \in \Delta = \left\{\delta : \delta \leq C_\delta \left(N/G\right)^{1/2}\right\}$ for some constant $C_\delta < \underline{\rho}\sqrt{\underline{kc}/2}$.

The range defined in Assumption 2 is wide. It accommodates first-stage testing procedures with a fixed nominal size. It also allows for testing procedures that adjust the critical value for an increasing number of first-stage tests. These include Bonferroni's correction and other more liberal rules for false discovery proportion or false discovery rate control under some additional mild rate conditions. See detailed discussions in Lemma A1 in the Appendix.

Lemma 1. *Let* $s_{sel,int} = \sigma_u/\sqrt{\sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 k_g p_g}$. *Suppose Assumptions 1 and 2 hold. Then, as* $G, N \to \infty$*:*

1. *If* $G^2/N \to 0$, *then* $\sqrt{N}(\hat\beta_{sel,int}(\delta) - \beta)/s_{sel,int} \Rightarrow N(0,1)$.
2. *If* $G/N \to 0$, *then* $\sqrt{N}(\hat\beta_{sssel,int}(\delta) - \beta)/s_{sel,int} \Rightarrow N(0,1)$.

The lemma has several interesting implications. First, unlike the select-and-pool method discussed in the previous section, the select-and-interact estimator has conventional large sample inference. Intuitively, this is because interacting the instrument with group indicators essentially re-weights the instrument by the estimated first-stage slope coefficient. This re-weighting changes the order of magnitude at which the exclusion restriction is violated through first-stage selection. At any finite sample, the exclusion restriction of the select-and-interact method is still violated, but the order of violation goes to zero faster than the local rate $1/\sqrt{N}$ and, therefore, does not have first-order impact on inference.

Under the assumptions of Lemma 1, $\hat\beta_{sel,int}(\delta)$ and $\hat\beta_{sssel,int}(\delta)$ are first-order asymptotically equivalent and efficient for all δ satisfying Assumption 2. This equivalence result, however, is not reflective of the finite-sample behavior of the two estimators. The weaker growth condition between G and N required in the second part of Lemma 1 suggests that the higher-order asymptotic bias and/or higher-order efficiency loss terms of the split-sample select-and-interact estimator might be of smaller order of magnitude than the full-sample select-and-interact estimator. Next, we formalize this argument by deriving the asymptotic MSEs of the two estimators as functions of δ.

3.2. Characterization of asymptotic mean squared error

To approximate the MSEs of the select-and-interacted 2SLS estimators, we apply the higher-order asymptotic expansion techniques in Nagar (1959), Donald and Newey (2001), Okui (2009), Cheng et al. (2019), and others. To keep the calculations tractable, we assume in this section that the error terms (u, v) follow a joint normal distribution. We write $\Phi(\cdot)$ and $\phi(\cdot)$ for the cumulative distribution function of the standard normal distribution and the probability density function of the standard normal distribution function, respectively.

Theorem 2. *Under Assumptions 1 and 2 and the additional assumption that* (u, v) *follow joint normal distribution, we have that*

1. *if* $G^2/N \to 0$ *as* $G, N \to \infty$, *the asymptotic MSE of* $\hat\beta_{sel,int}(\delta)$ *can be decomposed to*

$$
\begin{aligned}
&N(\hat\beta_{sel,int}(\delta) - \beta)^2 = \hat{Q}_{sel,int}(\delta) + \hat{r}_{sel,int}(\delta),\\
&E[\hat{Q}_{sel,int}(\delta)|\tilde{Z}, X] = \sigma_u^2/H + S_{sel,int}(\delta) + T_{sel,int}(\delta),\\
&\sup_{\delta\in\Delta}\left|(\hat{r}_{sel,int}(\delta) + T_{sel,int}(\delta))/S_{sel,int}(\delta)\right| = o_p(1),
\end{aligned}
$$

where $H = \frac{1}{N}\sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 Z_g' Z_g$ *and* $H^2 S_{sel,int}(\delta) = \sigma_{uv}^2 \left(\sum_g \left(1 - \Phi\left(\frac{\delta-\mu_g}{\sigma_v}\right) + \left(\frac{\delta}{\sigma_v}\right)\phi\left(\frac{\delta-\mu_g}{\sigma_v}\right)\right)\right)^2/N.$

2. *if* $G/N \to 0$ *and* $G_{+,w}/G \to 0$ *as* $G, N \to \infty$, *the asymptotic MSE of* $\hat\beta_{sssel,int}(\delta)$ *can be decomposed to*

$$
\begin{aligned}
&N(\hat\beta_{sssel,int}(\delta) - \beta)^2 = \hat{Q}_{sssel,int}(\delta) + \hat{r}_{sssel,int}(\delta),\\
&E[\hat{Q}_{sssel,int}(\delta)|\tilde{Z}, X] = \sigma_u^2/H + S_{sssel,int}(\delta) + T_{sssel,int}(\delta),\\
&\sup_{\delta\in\Delta}\left|(\hat{r}_{sssel,int}(\delta) + T_{sssel,int}(\delta))/S_{sssel,int}(\delta)\right| = o_p(1),
\end{aligned}
$$

where $H^2 S_{sssel,int}(\delta) = A_{sssel,int}(\delta) + B_{sssel,int}(\delta) + C_{sssel,int}(\delta)$ *with*

$$
\begin{aligned}
A_{sssel,int}(\delta) =& 2\sigma_u^2\sigma_v^2 \sum_g \left(1 - \Phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right) + \left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right)\phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right)\right)/N,\\
B_{sssel,int}(\delta) =& \sigma_u^2 \sum_g \mu_g^2 \Phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right)/N
\end{aligned}
$$

$$C_{sssel,int}(\delta) = 2\sigma_{uv}^2 \sum_g \left(1 - \Phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right) + \frac{\delta}{\sigma_v}\phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right)\right)^2 /N.$$

The first-order terms in the asymptotic MSE decomposition are expected given the asymptotic variance formula in Lemma 1. The higher-order terms of the full-sample select-and-interact estimator come from three different sources: many-IV bias, bias introduced by first-stage selection, and efficiency loss from falsely excluding groups with relevant instruments. Under the range for δ specified in Assumption 2, the higher-order efficiency loss term in the asymptotic MSE is dominated in order of magnitude by the bias terms. If first-stage selection was orthogonal to second-stage estimation, the bias term would be $\sigma_{uv}^2\left(\sum_g\left(1-\Phi\left(\frac{\delta-\mu_g}{\sigma_v}\right)\right)\right)^2/N$. The additional terms in $S_{sel,int}(\delta)$ represent extra asymptotic bias introduced from selection in the first-stage.

The repeated split-sample estimator $\hat{\beta}_{sssel,int}$ has the same first-order asymptotic MSE term as the full-sample select-and-interact estimator but its higher-order leading term is of a smaller order, or G/N. The term $A_{sssel,int}(\delta)$ represents the higher order bias of the split-sample estimators. The term $B_{sssel,int}(\delta)$ is a higher-order efficiency loss term due to excluding groups from the final estimator. The term $C_{sssel,int}(\delta)$ is a higher-order term due to combining the two split-sample estimators.

3.3. Adaptive δ selection for optimal MSE

In this section, we discuss how to select the thresholding value, δ, adaptively to achieve second-stage MSE-optimality relative to the expansion in Theorem 2.

Corollary 1. *Under Assumption 1 and the rate condition $G/N \to 0$ as $G, N \to \infty$, $\inf_\delta L(\delta) = 2b\sigma_u^2\sigma_v^2(1+\rho_{uv}^2)\frac{G}{N} + o_p(\frac{G}{N})$ if $G_{+,w}/G \to 0$, where b is defined in Assumption 1.5 and $L(\delta) = A_{sssel,int}(\delta) + B_{sssel,int}(\delta) + C_{sssel,int}(\delta)$.*

Corollary 1 establishes the optimal level of the asymptotic MSE of the repeated split-sample select-and-interact estimator when the proportion of groups with weak first-stage coefficients vanishes. In this case, the minimum asymptotic MSE is achieved when the thresholding value of δ singles out all groups with strong first-stage identification in the limit. When the proportion of groups with weak first-stage coefficients does not vanish, the minimum asymptotic MSE is still of order G/N, but the constant depends on the distribution of μ_g for $g \in \mathcal{G}_{+,w}$, often in a very complicated fashion. Moreover, it is not possible to consistently estimate the optimal constant, which is akin to the impossibility result for post-model selection estimators in Leeb and Pötscher (2005). As a result, characterizing the minimum asymptotic MSE level of $\hat{\beta}_{sssel,int}$ in the not well-separated case may not lead to a meaningful adaptive procedure for choosing the optimal threshold value δ given the data.

Following Corollary 1, let $L^* = 2b\sigma_u^2\sigma_v^2(1+\rho_{uv}^2)\frac{G}{N}$. The next theorem suggests an adaptive estimator for the optimal thresholding value δ whose leading higher order term in asymptotic MSE is equivalent to L^*. The theorem requires an additional assumption on the tail behavior of the instrument distribution and a slightly stronger rate condition that $G\log G/N \to 0$ as $G, N \to \infty$.

Theorem 3. *Let $(\hat{\sigma}_u^2, \hat{\sigma}_v^2, \hat{\sigma}_{uv}^2)$ be consistent estimators of $(\sigma_u^2, \sigma_v^2, \sigma_{uv}^2)$ and $\hat{\mu}_{(g)}$ be the order statistic such that $\hat{\mu}_{(1)} \geq \hat{\mu}_{(2)} \cdots \geq \hat{\mu}_{(G)}$. Let*

$$\hat{\mathcal{R}}(K) = \frac{\hat{\sigma}_u^2}{N}\sum_{g=K+1}^{G}\check{\mu}_{(g)}^2 + 2(\hat{\sigma}_u^2\hat{\sigma}_v^2 + \hat{\sigma}_{uv}^2)\frac{K}{N},$$

where $\check{\mu}_{(g)} = \hat{\mu}_{(g)}/\sqrt{\kappa_{G,N}}$, and $\kappa_{G,N}$ is a tuning sequence of order higher than $\log G$ and at most $\sqrt{\frac{N}{G}\log G}$ used to adjust for the first-stage estimation of ρ_g. Let $\hat{K} = argmin_K\hat{\mathcal{R}}(K)$ and $\hat{\delta} = \check{\mu}_{(\hat{K})}$. Under Assumption 1, $G\log G/N \to 0$ as $G, N \to \infty$, and the assumption that the instrument $\tilde{Z}$ follows a sub-exponential distribution, we have that

$$L(\hat{\delta})/L^* \xrightarrow{p} 1.$$

For $\hat{\delta}$ as in Theorem 3, we define the adaptive estimator $\hat{\beta}_{adpt} \equiv \hat{\beta}_{sssel,int}(\hat{\delta})$. Theorem 3 implies that for $\hat{\delta}$ equal to the $\hat{K}$th order statistics of $\check{\mu}$, the adaptive estimator has a leading higher-order asymptotic MSE term that converges to the minimum stated in Corollary 1. Note that the convergence result in this theorem does not require the proportion of groups with weak first-stage identification to vanish in the limit as assumed in Corollary 1. When $G_{+,w}/G \to 0$ holds, the adaptive estimator has optimal asymptotic MSE in both the first order and the leading higher order terms. When $G_{+,w}/G \to 0$ does not hold, the adaptive estimator is still first order efficient, but may not be higher order optimal among all split-sample select-and-interact estimators defined in Eq. (5).

The tuning parameter $\kappa_{G,N}$ is used as a wedge to separate the groups with strong first-stage signals from those with weak or irrelevant instruments when the first-stage parameter ρ_g is replaced by its estimator. Intuitively, $\kappa_{G,N}$ is chosen to dominate all $\hat{\mu}_g$ terms in $\mathcal{G}_{+,w}$ and $\mathcal{G}_0$ groups and be dominated by all $\hat{\mu}_g$ terms in $\mathcal{G}_{+,s}$ groups such that $\hat{\mathcal{R}}(.)$ is minimized at a value $\hat{\mathcal{R}}(\hat{K})$ that in the limit keeps all strong groups in the sample but discard all others. We set the rule-of-thumb $\kappa_{G,N}$ to $\kappa^*_{G,N} = (\log G)^2$ in the simulations and empirical sections. In the empirical section, we report robustness checks with alternative choices for $\kappa_{G,N}$ ($2\kappa^*_{G,N}$ and $\kappa^*_{G,N}/2$). We find the empirical results robust to such perturbations.

Under correct selection of the strong first-stage groups, the adaptive estimator is equivalent to the oracle estimator

$$\hat{\beta}_{oracle} = \Big(\sum_{g\in\mathcal{G}_{+,s}} \hat{\rho}_g Z_g' W_g\Big)^{-1} \sum_{g\in\mathcal{G}_{+,s}} \hat{\rho}_g Z_g' Y_g$$

that employs the identity of groups with a strong first-stage identification. Following the same arguments as in Section 2.2.2, it can be seen that $\hat{\beta}_{oracle}$ has probability limit $\sum_{g\in\mathcal{G}_{+,s}} \frac{\rho_g^2 V_g p_g}{\sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 V_g p_g}\beta_g$. Because correct selection of strong first-stage groups occurs with probability approaching one under the conditions in Theorem 3, the same weighted average causal effect interpretation of the oracle estimator is valid for the adaptive estimator, and

$$\hat{\beta}_{adpt} = \sum_{g\in\mathcal{G}_{+,s}} \frac{\rho_g^2 V_g p_g}{\sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 V_g p_g}\beta_g + o_p(1).$$

Our proposed adaptive procedure is akin to a version of the split-sample lasso selection estimator of Belloni et al. (2012). In simulations, we find that our proposed adaptive estimator behaves comparably and in some DGPs better than split-sample lasso in terms of MSE. In the two empirical applications, the two methods give similar point estimates and standard errors across almost all specifications, although their exact groups selected for 2SLS estimation often differ slightly.

4. Monte Carlo simulations

In this section, we study the finite-sample performance of different IV estimation procedures under first-stage heterogeneity. We use three data generating processes. Let $X_i, \tilde{Z}_i, v_i, e_i \sim i.i.d.\ N(0,1)$, and $u_i = \rho_{u,v} v_i + \sqrt{1-\rho_{u,v}^2}\, e_i$ with varying correlation coefficient $\rho_{u,v}$. We employ the simultaneous equation model in (1) with $\beta = 0$, and $\theta = \gamma = 1$ to generate the endogenous variables Y_{ig} and W_{ig}. The parameter ρ_g controls the relevance of instrument Z in group g and varies across DGPs.

Fig. 1 summarizes the distribution of ρ_g for the three DGPs in the simulations. We fix group size to $n_g = 500$ throughout. DGP 1 represents the case with well-separated first-stage signals. Out of G groups, where G varies from 40 to 200 in the simulations, a proportion p_s of them have strong first-stage ($\rho_g = 1$). For the rest of the groups the instrument is not correlated with the endogenous variable ($\rho_g = 0$). The first two graphs from the left in Fig. 1 plot the cumulative distribution functions (CDFs) of ρ_g in DGP 1 with $p_s = 0.25$ and $p_s = 0.05$, respectively. In DGP 2, we mix in some non-negligible proportion, p_w, of weak groups where $\rho_g = 0.2$. The third and fourth graphs from the left in Fig. 1 plot the CDFs of DGP 2, where $p_s = p_w = 0.125$ and $p_s = p_w = 0.025$, respectively. The last graph plots DGP 3, which represents a case where the weak and strong groups have no separation. Ninety percent of the groups in DGP 3 have irrelevant instruments. Among the remaining ten percent of groups, half of them have first-stage effect $\rho_g \sim \mathcal{N}(0.2, 0.1^2)$ and the other half have $\rho_g \sim \mathcal{N}(1, 0.25^2)$. Motivated by the data patterns of the two empirical examples in Section 5, all DGPs considered in this section have large proportions of groups with zero first-stage coefficients.

In our simulations, we study the performance of the following estimators: (1) $\hat{\beta}_{pool}$ (2SLS-P) the conventional pooled 2SLS estimator that ignores first-stage heterogeneity, (2) $\hat{\beta}_{int}$ (2SLS-INT) the 2SLS that uses full interaction of the scalar instrumental variable with all group dummies as the instruments, (3) the repeated split-sample version of 2SLS-INT, denoted as 2SLS-SSINT, (4) the infeasible repeated split-sample interacted 2SLS that chooses the thresholding value δ to minimize the theoretical MSE of the split-sample select-and-interact estimator stated in Theorem 2 using oracle information of ρ_g, (5) the limited information maximum likelihood estimator (LIML-INT) with interact the instrument with all group dummies. We also consider two recent variants of the JIVE estimator (Angrist et al., 1999): the UJIVE estimator proposed by Kolesár (2013) and (7) the improved JIVE (IJIVE) estimator proposed by Ackerberg and Devereux (2009) to address the many-covariates bias of JIVE. Lastly we consider (8) a split-sample 2SLS estimator that uses lasso for first-stage selection (2SLS-SSL) among fully interacted instruments, and (9) our proposed split-sample adaptive estimator (2SLS-ADPT) with the thresholding value estimated from the data using Theorem 3 and $\kappa_{G,N} = (\log(G))^2$. 2SLS-INT and 2SLS-SSINT correspond to $\hat{\beta}_{sel,int}(-\infty)$ and $\hat{\beta}_{sssel,int}(-\infty)$ defined in Section 3.1, respectively.

Tables 2–4 report empirical MSE (except for LIML-INT, which does not have moments) and MAD (median absolute deviation) across 500 simulations, as well as rejection rates for the second-stage t-test for the three DGPs under two different error distributions: normal and chi-squared with 3 degrees of freedom (χ_3^2). Notice that our paper focuses on asymptotic MSE expansion and optimality. MAD and t-test comparisons are reported as robustness checks.

For all DGPs, the pooled two stage least square estimator (2SLS-P) has very poor MSE and MAD performance. This is mainly driven by variance inflation: the large number of groups with a zero first-stage effect makes the pooled 2SLS estimator inefficient. 2SLS-INT is first-order efficient. It behaves well when the number of instruments (i.e., G in our setup) is small and the first-stage signal (i.e., the proportion of nonzero groups in our setup) is strong. However, because

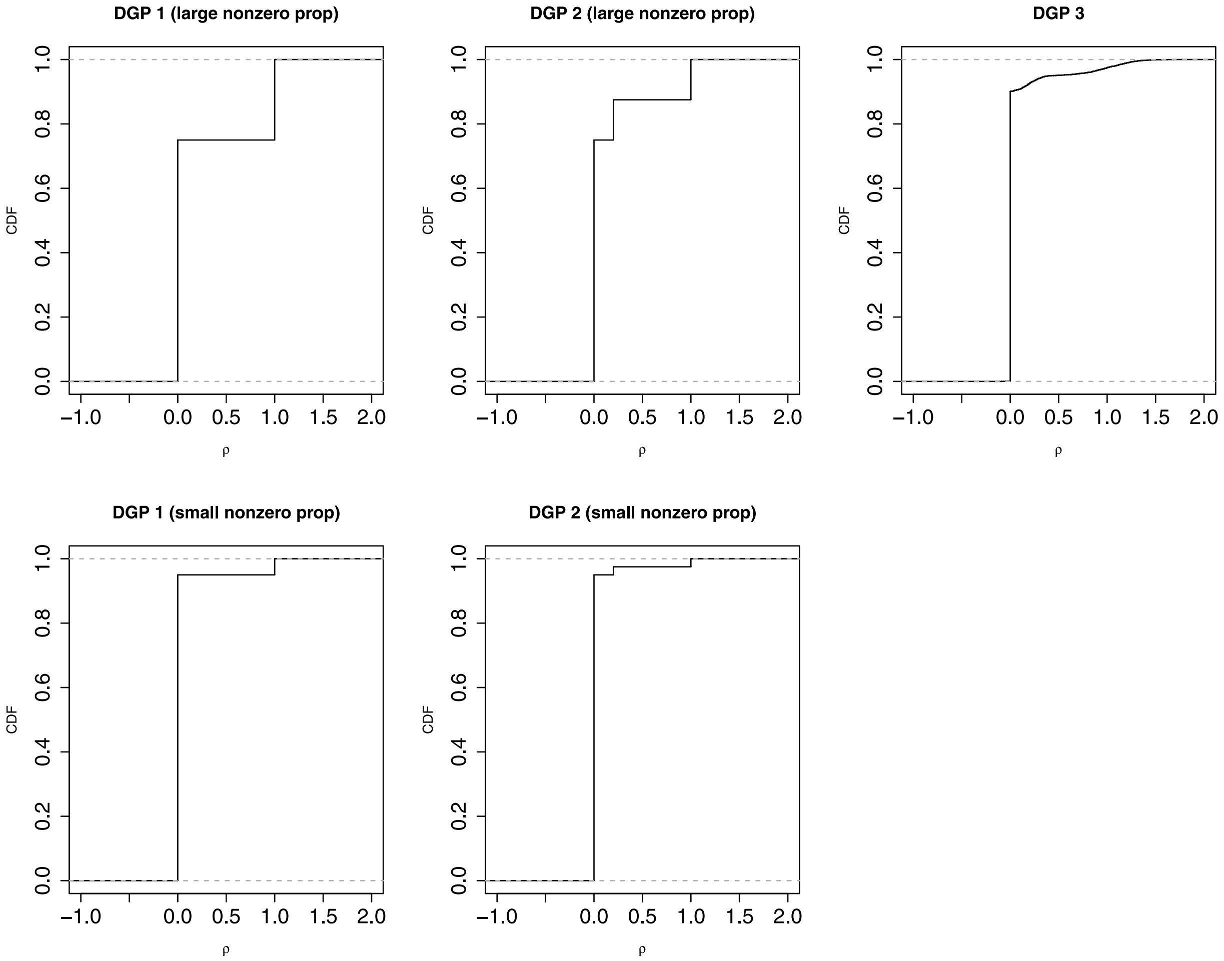

Fig. 1. Distribution of ρ_g for three DGPs.

of many-IV bias, 2SLS-INT can have much higher MSE and MAD than 2SLS-INF when G is large and the proportion of nonzero group is small.[4] In such cases, 2SLS-SSINT helps reduce the asymptotic bias, but can still be improved in terms of asymptotic MSE. These observations provide motivation for our proposed estimator, which uses a data-driven procedure to mimic 2SLS-INF. We next study the performance of 2SLS-ADPT, 2SLS-SSL, LIML-INT, UJIVE and IJIVE relative to the infeasible estimator 2SLS-INF.

For DGP1 with normal errors, the proposed adaptive estimator (2SLS-ADPT) is almost as good as the corresponding infeasible estimator 2SLS-INF. All the other competitors: 2SLS-SSL, LIML-INT, IJIVE and UJIVE also behave similarly well. LIML-INT is a very competitive estimator (in terms of MAD) under normal error which is not surprising given its optimality (Kolesár, 2018). However, as discussed in Kolesár (2013), when treatment effects vary across groups LIML-INT does not have a clear causal interpretation.

For the same DGP but with a χ^2_3 error distribution and a small proportion of non-zero groups, there is a noticeable gap between the MSE of the 2SLS-INF and that of the 2SLS-ADPT estimator. However, 2SLS-ADPT outperforms LIML-INT, 2SLS-SSL, IJIVE and UJIVE when the proportion of non-zero first-stage signal is small ($p_s = 0.05$). When $p_s = 0.25$, they perform similarly. With a χ^2_3 error, LIML-INT has noticeable size inflation. This behavior of LIML under non-normal errors has been previously documented in Hahn et al. (2004) and Sølvsten (2020), among others.

Table 3 reports simulation results for DGP2, which adds a small proportion of weak first-stage groups. For DGP2, both 2SLS-ADPT and 2SLS-SSL, which are designed to learn the first-stage identification structure, out-perform the 2SLS-P, 2SLS-INT and 2SLS-SSINT estimators. Among the competitors of 2SLS-ADPT and 2SLS-SSL (e.g. LIML-INT, IJIVE and UJIVE), when p_s and p_w equals to 0.025 respectively, our proposed estimator is the best in terms of MSE, and when p_s and p_w

[4] In some settings, especially when the proportion of strong groups is large, 2SLS-INT can perform similarly to 2SLS-INF. For example, under DGP1 with $p_s = 0.5$ (not reported), 2SLS-INT has similar MSE and MAD as 2SLS-INF. Motivated by the empirical applications in Section 5, the simulation designs consider settings with a small or moderate fraction of strong groups.

Table 2
Rejection rates and MSE performance for DGP 1.

		2SLS-P	2SLS-INT	2SLS-SSINT	2SLS-INF	2SLS-ADPT	LIML-INT	2SLS-SSL	UJIVE	IJIVE
$p_s = 0.05$, $\rho_{u,v} = 0.25$ and normal errors										
G = 40	$N \times$ MSE	436.882	21.215	22.669	20.390	20.390	–	21.579	21.144	21.198
	$N \times$ MAD	1923.779	443.467	423.261	430.740	430.740	432.387	426.859	420.578	420.594
	Rej. Prop.	0.038	0.066	0.046	0.050	0.050	0.058	0.050	0.048	0.048
G = 100	$N \times$ MSE	458.475	23.663	20.486	19.609	19.599	–	19.808	19.972	19.950
	$N \times$ MAD	3318.415	761.088	646.376	627.586	627.586	593.404	625.040	632.902	632.412
	Rej. Prop.	0.058	0.082	0.038	0.048	0.048	0.052	0.048	0.032	0.032
G = 200	$N \times$ MSE	387.056	32.099	23.843	21.686	21.683	–	22.479	22.392	22.414
	$N \times$ MAD	4341.239	1267.731	1043.499	1021.058	1018.517	1001.641	1007.512	1020.777	1025.588
	Rej. Prop.	0.038	0.150	0.058	0.056	0.056	0.068	0.064	0.022	0.026
$p_s = 0.25$, $\rho_{u,v} = 0.25$ and normal errors										
G = 40	$N \times$ MSE	16.240	4.221	4.334	4.215	4.215	–	4.253	4.276	4.277
	$N \times$ MAD	385.748	206.675	202.563	198.870	198.870	203.090	196.234	204.595	203.852
	Rej. Prop.	0.054	0.042	0.042	0.042	0.042	0.048	0.042	0.050	0.054
G = 100	$N \times$ MSE	17.488	4.594	4.390	4.399	4.399	–	4.374	4.391	4.390
	$N \times$ MAD	661.972	311.558	306.842	320.676	320.676	307.353	322.333	312.232	312.709
	Rej. Prop.	0.062	0.076	0.066	0.072	0.072	0.078	0.068	0.066	0.068
G = 200	$N \times$ MSE	15.208	4.578	4.156	4.051	4.050	–	4.093	4.079	4.079
	$N \times$ MAD	875.906	469.002	427.221	421.565	421.565	418.238	425.306	423.464	421.793
	Rej. Prop.	0.042	0.068	0.056	0.050	0.050	0.058	0.052	0.056	0.048
$p_s = 0.05$, $\rho_{u,v} = 0.25$ and χ^2_3 errors										
G = 40	$N \times$ MSE	6113.771	135.939	199.797	129.725	136.969	–	153.161	167.058	166.815
	$N \times$ MAD	5278.398	1145.955	1321.611	1048.631	1052.631	1127.648	1156.241	1136.268	1151.079
	Rej. Prop.	0.028	0.108	0.042	0.054	0.056	0.082	0.058	0.036	0.036
G = 100	$N \times$ MSE	2791.348	189.829	191.282	124.294	142.841	–	143.370	150.963	151.634
	$N \times$ MAD	7135.788	2340.853	1928.876	1541.093	1632.401	1796.948	1658.365	1753.012	1775.675
	Rej. Prop.	0.026	0.174	0.054	0.058	0.064	0.074	0.054	0.028	0.024
G = 200	$N \times$ MSE	2663.635	314.390	183.905	121.197	132.861	–	137.191	150.592	150.432
	$N \times$ MAD	10 743.992	4569.327	2901.436	2378.734	2545.821	2497.460	2494.414	2495.732	2486.262
	Rej. Prop.	0.040	0.302	0.052	0.040	0.038	0.068	0.046	0.008	0.010
$p_s = 0.25$, $\rho_{u,v} = 0.25$ and χ^2_3 errors										
G = 40	$N \times$ MSE	106.285	25.804	26.797	25.590	25.747	–	26.262	26.065	26.119
	$N \times$ MAD	1072.737	492.951	532.019	496.556	501.732	486.464	532.610	500.576	504.090
	Rej. Prop.	0.050	0.068	0.042	0.050	0.054	0.060	0.054	0.040	0.032
G = 100	$N \times$ MSE	94.806	30.146	29.370	26.310	27.498	–	27.206	27.098	27.113
	$N \times$ MAD	1439.007	799.533	803.419	813.880	790.668	792.318	776.718	772.060	762.407
	Rej. Prop.	0.040	0.080	0.058	0.058	0.060	0.064	0.060	0.038	0.044
G = 200	$N \times$ MSE	98.786	34.913	27.915	25.958	26.671	–	26.435	26.682	26.655
	$N \times$ MAD	2165.718	1327.275	1159.424	1110.788	1140.030	1123.004	1140.352	1096.351	1097.774
	Rej. Prop.	0.048	0.112	0.048	0.046	0.048	0.046	0.046	0.016	0.026

Note: DGP1 under normal and χ^2_3 errors. Scaled mean squared error, absolute sum of error and rejection probability are reported for different configurations of G, p_s, and p_w. The group sample size is fixed at $n_g = 500$. Results are based on 500 simulation repetitions.

increases to 0.125 respectively, IJIVE or UJIVE out-perform slightly in terms of MSE. In terms of MAD, there is no a clear top performer.

Table 4 reports simulation results for DGP3, which features groups with weak and strong first-stage effects that are not well-separated. For DGP3, both 2SLS-ADPT and 2SLS-SSL out-perform the 2SLS-P, 2SLS-INT and 2SLS-SSINT estimators in almost all settings (i.e. various G and two different error distributions). Comparing to LIML-INT, UJIVE and IJIVE, our proposed 2SLS-ADPT estimator outperforms in terms of MSE or MAD when the error distribution is χ^2_3.

5. Empirical examples

5.1. Return to compulsory schooling

The return to schooling literature studies how an extra year of schooling affects individual outcomes later in life, such as earnings and health outcomes. Years of schooling may correlate with omitted variables, such as early cognitive ability and family background. For this reason, researchers often use variation in compulsory schooling laws across states and across time in the U.S. (see, Lleras-Muney, 2005, Oreopoulos, 2006, and Stephens and Yang, 2014, among others) and other countries (Oreopoulos, 2006) to instrument for years of schooling. The argument for identification is that any law

Table 3
Rejection rates and MSE performance for DGP 2.

		2SLS-P	2SLS-INT	2SLS-SSINT	2SLS-INF	2SLS-ADPT	LIML-INT	2SLS-SSL	UJIVE	IJIVE
$p_s = p_w = 0.025$, $\rho_{u,v} = 0.25$ and normal errors										
G = 40	N × MSE	1435.660	43.079	47.703	40.468	41.944	–	44.084	43.254	43.411
	N × MAD	3239.714	621.016	661.890	610.296	623.647	603.854	615.128	615.956	618.207
	Rej. Prop.	0.026	0.076	0.058	0.048	0.048	0.064	0.056	0.046	0.042
G = 100	N × MSE	1905.386	64.325	54.582	46.870	48.225	–	50.044	50.502	50.486
	N × MAD	6199.708	1318.320	1069.955	1036.453	1017.512	1016.949	1041.127	1007.841	1018.069
	Rej. Prop.	0.038	0.118	0.050	0.046	0.042	0.070	0.056	0.022	0.018
G = 200	N × MSE	1120.562	75.823	47.525	40.191	40.989	–	42.847	42.896	42.950
	N × MAD	7228.982	2043.003	1493.240	1459.994	1445.584	1477.341	1478.735	1441.131	1445.072
	Rej. Prop.	0.032	0.202	0.064	0.054	0.050	0.066	0.060	0.010	0.014
$p_s = p_w = 0.125$, $\rho_{u,v} = 0.25$ and normal errors										
G = 40	N × MSE	45.407	8.481	8.708	8.414	8.546	–	8.575	8.421	8.428
	N × MAD	647.919	265.315	276.036	260.195	272.046	261.139	270.705	262.210	264.634
	Rej. Prop.	0.054	0.076	0.062	0.054	0.056	0.050	0.058	0.058	0.050
G = 100	N × MSE	51.337	9.325	8.816	8.750	9.021	–	8.706	8.680	8.680
	N × MAD	1145.194	464.062	431.866	431.265	426.228	428.505	429.941	438.845	442.226
	Rej. Prop.	0.062	0.062	0.064	0.058	0.066	0.070	0.060	0.056	0.058
G = 200	N × MSE	42.263	9.673	8.297	7.873	8.114	–	8.003	8.049	8.047
	N × MAD	1456.95	673.573	604.213	582.558	565.383	570.043	583.826	588.996	583.131
	Rej. Prop.	0.040	0.080	0.050	0.054	0.052	0.046	0.050	0.044	0.042
$p_s = p_w = 0.025$, $\rho_{u,v} = 0.25$ and χ^2_3 errors										
G = 40	N × MSE	720 188.116	260.451	564.523	246.761	259.075	–	342.620	387.119	386.144
	N × MAD	8722.373	1709.158	2091.322	1523.892	1644.423	1703.432	1842.463	1680.142	1618.648
	Rej. Prop.	0.014	0.142	0.026	0.032	0.032	0.082	0.026	0.032	0.044
G = 100	N × MSE	98 979 469.114	543.929	729.808	293.429	369.645	–	399.998	448.403	451.981
	N × MAD	14 080.389	4194.958	3740.958	2470.492	2619.759	2883.843	2789.741	2825.120	2850.834
	Rej. Prop.	0.016	0.296	0.054	0.050	0.048	0.100	0.042	0.038	0.034
G = 200	N × MSE	9393.866	741.439	476.687	223.481	271.115	–	288.774	327.028	326.492
	N × MAD	18 096.993	7615.012	4307.353	3019.094	3377.892	3582.159	3350.654	3750.788	3771.590
	Rej. Prop.	0.036	0.482	0.052	0.034	0.044	0.094	0.040	0.012	0.018
$p_s = p_w = 0.125$, $\rho_{u,v} = 0.25$ and χ^2_3 errors										
G = 40	N × MSE	303.819	51.212	59.449	52.709	54.365	–	55.503	54.185	54.311
	N × MAD	1761.756	716.053	754.384	710.452	697.723	731.689	699.147	716.633	701.912
	Rej. Prop.	0.048	0.080	0.042	0.048	0.054	0.062	0.058	0.038	0.048
G = 100	N × MSE	279.308	68.797	67.898	57.511	60.779	–	60.541	59.851	60.007
	N × MAD	2458.124	1280.568	1188.038	1107.445	1163.068	1127.071	1093.128	1112.960	1118.185
	Rej. Prop.	0.040	0.116	0.068	0.064	0.078	0.078	0.072	0.028	0.042
G = 200	N × MSE	275.248	84.233	59.538	51.447	54.105	–	53.655	54.307	54.247
	N × MAD	3568.601	2127.199	1583.559	1398.683	1518.890	1544.208	1528.103	1517.198	1474.994
	Rej. Prop.	0.048	0.178	0.060	0.050	0.056	0.062	0.058	0.008	0.008

Note: DGP2 under normal and χ^2_3 errors. Scaled mean squared error, absolute sum of error and rejection probability are reported for different configurations of G, p_s, and p_w. The group sample size is fixed at $n_g = 500$. Results are based on 500 simulation repetitions.

change in minimum school leaving age may affect individual education attainment, but not individual well-being later in life, other than through the education channel.

In this section we re-analyze the public-use U.S. Census dataset compiled by Stephens and Yang (2014). In contrast to Stephens and Yang (2014), we explicitly model first-stage heterogeneity in the effects of compulsory schooling laws. Our first-stage regression interacts years of compulsory schooling as well as other exogenous controls with indicators for geographic regions and demographic groups. The dataset includes native-born individuals between 25 and 54 years of age across the 1960–1980 U.S. Decennial Censuses. We use subscripts i, t, and s, to index individuals, cohorts, and birth states, respectively. We consider the model

$$\begin{aligned} Log\,wage_{ist} &= \beta Educ_{ist} + \sum_{g=1}^{G} 1(S_{is} = g) X_{ist}\theta_g + u_{ist} \\ Educ_{ist} &= \sum_{g=1}^{G} \rho_g 1(S_{is} = g) CL_{st} + \sum_{g=1}^{G} 1(S_{is} = g) X_{ist}\gamma_g + v_{ist}, \end{aligned} \tag{6}$$

Table 4
Rejection rates and MSE performance for DGP 3.

		2SLS-P	2SLS-INT	2SLS-SSINT	2SLS-INF	2SLS-ADPT	LIML-INT	2SLS-SSL	UJIVE	IJIVE
DGP3 with $\rho_{u,v} = 0.25$ and normal errors										
G = 40	$N \times$ MSE	467.057	35.303	38.713	33.382	33.810	–	35.869	35.326	35.437
	$N \times$ MAD	1992.620	581.345	552.848	528.262	556.787	530.137	562.163	548.763	547.366
	Rej. Prop.	0.040	0.074	0.064	0.046	0.052	0.062	0.058	0.046	0.052
G = 100	$N \times$ MSE	390.835	25.589	21.891	20.821	21.321	–	20.999	21.314	21.296
	$N \times$ MAD	3071.019	786.978	686.456	648.956	668.154	632.800	671.309	661.978	661.012
	Rej. Prop.	0.058	0.084	0.046	0.054	0.054	0.052	0.052	0.038	0.028
G = 200	$N \times$ MSE	316.284	33.967	24.779	22.619	23.071	–	23.374	23.323	23.340
	$N \times$ MAD	3923.494	1328.891	1061.998	1039.967	1033.464	1061.145	1041.426	1051.217	1052.146
	Rej. Prop.	0.038	0.146	0.068	0.052	0.056	0.060	0.048	0.018	0.030
DGP3 with $\rho_{u,v} = 0.25$ and χ^2_3 errors										
G = 40	$N \times$ MSE	9623.419	206.304	398.164	213.024	260.572	–	290.787	294.854	294.622
	$N \times$ MAD	5459.984	1440.192	1897.443	1441.298	1512.635	1517.961	1627.616	1531.558	1503.804
	Rej. Prop.	0.026	0.134	0.028	0.042	0.050	0.086	0.040	0.030	0.034
G = 100	$N \times$ MSE	2318.217	201.848	204.249	132.450	158.688	–	158.259	162.105	162.755
	$N \times$ MAD	6620.035	2313.303	1980.954	1658.628	1879.294	1876.630	1761.201	1858.115	1868.637
	Rej. Prop.	0.026	0.186	0.052	0.048	0.056	0.084	0.054	0.016	0.018
G = 200	$N \times$ MSE	2150.827	327.106	190.661	125.153	138.088	–	143.897	155.546	155.411
	$N \times$ MAD	9760.407	4617.979	2970.616	2480.836	2663.850	2604.448	2496.421	2721.165	2706.670
	Rej. Prop.	0.042	0.308	0.054	0.030	0.030	0.080	0.050	0.010	0.010

Note: DGP3 under normal and χ^2_3 errors. Scaled mean squared error, absolute sum of error and rejection probability are reported for different configurations of G. The group sample size is fixed at $n_g = 500$. Results are based on 500 simulation repetitions.

where $Logwage_{ist}$ and $Educ_{ist}$ are the log wage and years of schooling of individual i, CL_{st} is the number of years of compulsory schooling that cohort t in state s faces at age 14, and S_{is} is a group indicator that varies with individual demographic characteristics and birth state. The exogenous regressor X_{ist} includes survey year, birth state, census division by gender and race, census division by birth year fixed effects, and a fourth-order polynomial in age (except for those specification where we use survey years to define groups; see below).

We use four definitions of groups to characterize heterogeneity in the first-stage correlation between the instrument and the endogenous regressor across groups: (A) census region by demographic control, (B) census division by demographic control, (C) census region by demographic control by survey year, and (D) census division by demographic control by survey year. The demographic control is a categorical variable with four categories: White males, White females, non-White males, and non-White females. Because non-White minorities only consist of 11.75 percent of the data sample (10.88 percent black, 0.87 percent other race), we pool non-White males and females together in a robustness check.

Besides allowing for first-stage heterogeneity, our simultaneous equation model is the same as the one in Table 1 of Stephens and Yang (2014), except that Stephens and Yang (2014) use three indicators (corresponding to being required to attend seven, eight, and nine or more years of schooling) constructed from CL_{st} as instruments, while we use CL_{st} directly. We adopt this specification because our formal results apply only to the scalar IV case. In addition, our specification includes census division by year-of-birth fixed effects, which is more robust than the census region by year-of-birth fixed effects in Stephens and Yang (2014). We adopt this specification because our group definitions (B) and (D) use census division to form groups. Recall that the method we propose allows for group-specific slopes for exogenous regressors including intercepts. To ensure the same set of fixed effect controls are used across regressions with all four group definitions, we upgrade the census region by birth year fixed effects used in Stephens and Yang (2014) to census division by birth year fixed effects.

Geographic groups are natural in our context because of heterogeneity in the enforcement of compulsory schooling laws and in school quality across the U.S. In addition, the effect of compulsory schooling laws on schooling depends on gender and race. Race has been found to affect the strength of the first-stage relationship in (6) (e.g., Lleras-Muney (2002) footnote 44). Gender is often used to define subsamples in studies on the effect of compulsory schooling. For example, Oreopoulos (2006) uses males and non-White males for subsample analysis, while the estimates in Stephens and Yang (2014) are for Whites and White males. Finally, we consider first-stage heterogeneity across survey years because of concerns about survey accuracy in the early years of our sample. The specifications that use survey years to define groups omit the fourth-order polynomial in age from X_{ist} (because age is then perfectly collinear with birth year fixed effects in groupwise regressions).

Fig. 2 plots p-values of groupwise first-stage upper one-sided t-tests. The first panel is for group definition C and the second panel is for group definition D. Both graphs show strong evidence of a mixture between groups with strong and weak/irrelevant first-stages. The graphs also show that some groups actually have a negative and statistically significant first-stage relationship between years of compulsory schooling and years of actual schooling (a p-value close to one for an upper one-sided t-test implies the rejection of the corresponding lower one-sided t-test with high confidence). This

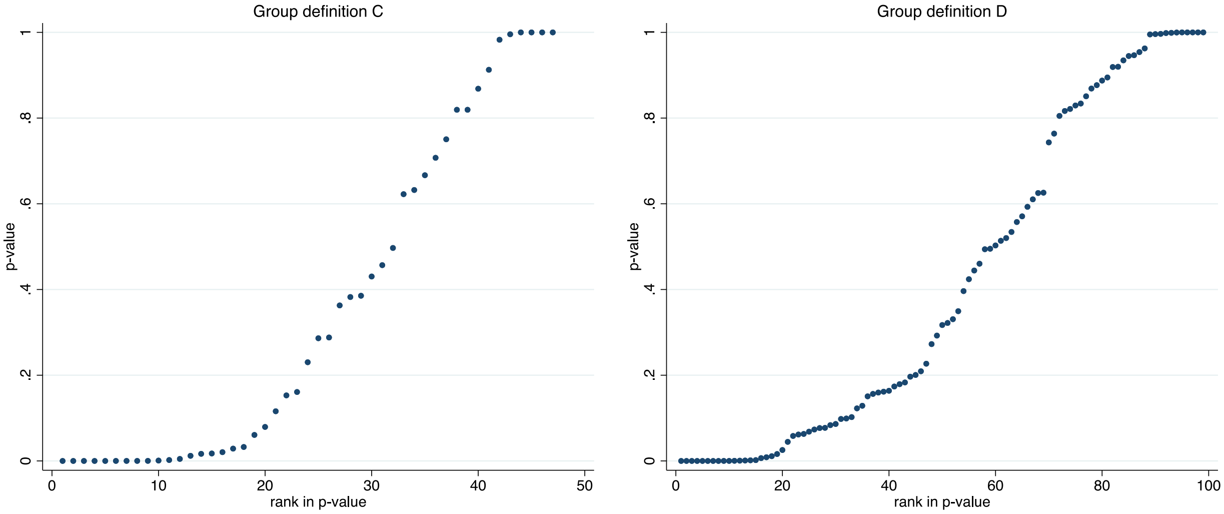

Fig. 2. Return to compulsory schooling: First-stage signal by groups. Note: Dataset is from Stephens and Yang (2014). The endogenous regressor is years of schooling. The instrument is the compulsory schooling year a birth cohort faces at age 14. All regressions also control state, survey year, subregion by birth year, gender, and race fixed effects. The graphs plot the top ten groupwise $\hat{\mu}_g$ against their corresponding first-stage $\hat{\rho}_g$ slope estimates.

could be the result of unrelated changes in the distribution of the variables that happen at the same time as changes in compulsory schooling, creating a threat to the validity of the exclusion restriction. By design, our adaptive procedure only selects groups with a strong and positive first-stage, which is also necessary for a LATE-type interpretation of 2SLS.

Table 5 reports regression results from various existing and proposed estimation methods. Panels A1–D1 use four gender and race categories, White males, White females, non-White males, and non-White females, to define groups. Panels A2–D2 use White males, White females, and non-White as a robustness check. Columns (1)–(5) report estimates from OLS, pooled 2SLS (2SLS-P), fully-interacted 2SLS (2SLS-INT), fully-interacted LIML (LIML-INT), and interacted 2SLS with repeated split-sample lasso selection of strong groups (2SLS-SSL), respectively. Columns (6)–(8) report estimation results from the proposed procedure, which is repeated split-sample 2SLS with adaptive selection of strong groups to minimize asymptotic MSE. Column (6) uses the tuning sequence $\kappa^*_{G,N} = (\log(G))^2$ discussed in Section 3. Columns (7) and (8) provide robustness checks of the proposed method using $2\kappa^*_{G,N}$ and $\kappa^*_{G,N}/2$, respectively. LIML-INT results are not reported in panels A1–B1 and A2–B2 because both *Stata* and *R* fail to compute LIML-INT for these specifications because of multicollinearity across census division by birth year fixed effects, census year indicators, and a fourth-order polynomial in age. LIML-INT results are reported in panels C1–D1 and C2–D2, which omit the fourth-order polynomial in age because of perfect collinearity with birth year in groupwise regressions. The 2SLS estimator in column (5) uses only lasso-selected groups with a positive first-stage relationship. The IV lasso computation is carried out with the default setting in *R* package *hdm* (Chernozhukov et al., 2016).

Stephens and Yang (2014) find that allowing for region by year-of-birth fixed effects often yields insignificant estimates of the return to compulsory schooling. This corresponds to the insignificant pooled 2SLS estimates across all eight rows of Table 5. All estimators reported in columns (3)–(8) are first-order equivalent under assumptions discussed in Section 3. When higher-order asymptotic MSE terms are considered, 2SLS-INT has the smallest asymptotic variance but could potentially suffer from nontrivial many-IV bias as shown in the simulations of Section 4. The proposed adaptive method, on the other hand, has better rates of higher-order asymptotic bias. As is seen from the table, 2SLS-INT has a small advantage in standard error relative to competing estimators. But it also has a larger point estimate, likely because many-IV bias makes the 2SLS-INT estimate close to the OLS estimate. Estimation results for the proposed adaptive procedure in columns (6)–(8) are mostly statistically significant but qualitatively smaller than both OLS in column (1) and 2SLS-INT in column (3). The estimates are also robust to perturbations in the definition of the tuning parameter sequence $\kappa_{G,N}$. In this application, 2SLS-SSL in column (5) gives similar point estimates and standard errors as the proposed procedure, although the exact groups selected for 2SLS regression are slightly different for different methods.

The graphs in Fig. 3 plot the groups with highest values of $\hat{\mu}_g$ for each specification to illustrate how the adaptive procedure selects groups in this empirical application. The top two graphs are for Panel A1 and A2 in Table 5, which correspond to the coarsest definition of groups. The bottom two graphs are for Panel D1 and D2, which correspond to the finest definition of groups. Each dot in the graphs represents a group. We color-code selection of two, one, or none of the sample splits by our adaptive 2SLS procedure. The results in the figure show that White males and White females in some divisions in the Northeast, Midwest, and South have larger contributions to first-stage identification than the rest. Non-White groups and groups in West divisions do not seem to contribute much to identification.

Table 5
Return to compulsory schooling: Estimation results.

Full-sample				Select-and-interact			
OLS	2SLS-P	2SLS-INT	LIML-INT	2SLS-SSL	2SLS-ADPT		
					(κ^*)	($2\kappa^*$)	($\kappa^*/2$)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panels A1–D1: four gender and race categories							
Panel A1: groups defined by census region, gender, and race							
0.070^{***}	−0.254	0.066^{***}	–	0.040^{***}	0.040^{***}	0.040^{***}	0.040^{***}
(0.000)	(0.145)	(0.011)	–	(0.015)	(0.015)	(0.015)	(0.015)
Panel B1: groups defined by census division, gender, and race							
0.070^{***}	−0.255	0.061^{***}	–	0.035^{***}	0.036^{***}	0.037^{***}	0.036^{***}
(0.000)	(0.145)	(0.009)	–	(0.013)	(0.013)	(0.013)	(0.013)
Panel C1: groups defined by census region, gender, race, and survey year							
0.070^{***}	−0.266	0.069^{***}	0.068^{***}	0.011	0.025	0.026	0.025
(0.000)	(0.170)	(0.013)	(0.015)	(0.022)	(0.021)	(0.021)	(0.021)
Panel D1: groups defined by census division, gender, race, and survey year							
0.070^{***}	−0.266	0.065^{***}	0.063^{***}	0.036^{**}	0.0036^{***}	0.0037^{***}	0.036^{**}
(0.000)	(0.170)	(0.009)	(0.010)	(0.015)	(0.014)	(0.014)	(0.014)
Panels A2–D2: three gender and race categories							
Panel A2: groups defined by census region, gender, and race							
0.069^{***}	−0.243	0.063^{***}	–	0.041^{***}	0.041^{***}	0.041^{***}	0.041^{***}
(0.000)	(0.155)	(0.011)	–	(0.015)	(0.015)	(0.015)	(0.015)
Panel B2: groups defined by census division, gender, and race							
0.069^{***}	−0.242	0.057^{***}	–	0.038^{***}	0.037^{***}	0.038^{***}	0.037^{***}
(0.000)	(0.154)	(0.009)	–	(0.013)	(0.012)	(0.013)	(0.013)
Panel C2: groups defined by census region, gender, race, and survey year							
0.069^{***}	−0.257	0.066^{***}	0.065^{***}	0.011	0.024	0.021	0.024
(0.000)	(0.174)	(0.013)	(0.015)	(0.022)	(0.021)	(0.021)	(0.021)
Panel D2: groups defined by census division, gender, race, and survey year							
0.069^{***}	−0.257	0.063^{***}	0.061^{***}	0.036^{**}	0.035^{**}	0.036^{**}	0.034^{**}
(0.000)	(0.174)	(0.009)	(0.010)	(0.015)	(0.015)	(0.015)	(0.015)

Note: Dataset is from Stephens and Yang (2014). The endogenous regressor is years of schooling. The instrument is the compulsory schooling year a birth cohort faces at age 14. All regressions also control state, survey year, census division by birth year, and census division by gender and race fixed effects. Panels A1–D1 use four demographic groups: White males, White females, non-White males, and non-White females. Panels A2–D2 use three demographic groups as a robustness check: White males, White females, and the non-Whites. Regressions in Panels A1–B1 and A2–B2 also include a fourth-order polynomial in age.

5.2. *Voter turnout*

Charles and Stephens (2013) uses county-level data to study the effect of local labor market variables, such as wages or employment rates, on voter turnout in various U.S. elections, including elections for governor, senator, US Congress, state House of Representatives, and U.S. President. The identification strategy first differences out county-level fixed effects and then accounts for potentially endogenous changes in local market activities using exogenous shocks to oil/natural gas (oil, hereafter) and coal supply. This strategy follows the earlier work by Black et al. (2002), which utilizes coal shocks to study the impact of local economic conditions on participation in programs of disability payments, and Acemoglu et al. (2013), which utilizes oil shocks to study the effect of local income on health spending. Recently, Charles et al. (2018) also uses oil shocks to study the effect of local labor market conditions on disability take-up in federal programs.

The articles mentioned above measure energy shocks as changes in national employment in energy production industries or global energy price, interacted with a measure of the importance of energy industry in a county prior to the period of study. The identification power of the instrument varies across states, and the authors in this literature often restrict the sample to a pre-selected list of oil and/or coal states. For example, Charles and Stephens (2013) defines coal states as Kentucky, Ohio, Pennsylvania, and West Virginia, following Black et al. (2002), and defines oil states as Colorado, Kansas, Mississippi, Montana, New Mexico, North Dakota, Oklahoma, Texas, Utah, and Wyoming, those with at least 1 percent of annual state wages in the 1974 County Business Patterns (CBP) in the oil industry. Charles et al. (2018) adds Louisiana to the list of oil states. Acemoglu et al. (2013) uses a sample of southern states.

In this section, we revisit Charles and Stephens (2013). We adopt the same model specification as in Charles and Stephens (2013), except for a modification in the definition of the instrumental variable, as explained below. Also, instead of using a pre-determined list of oil and coal states, we select states for our sample using our proposed adaptive procedure.

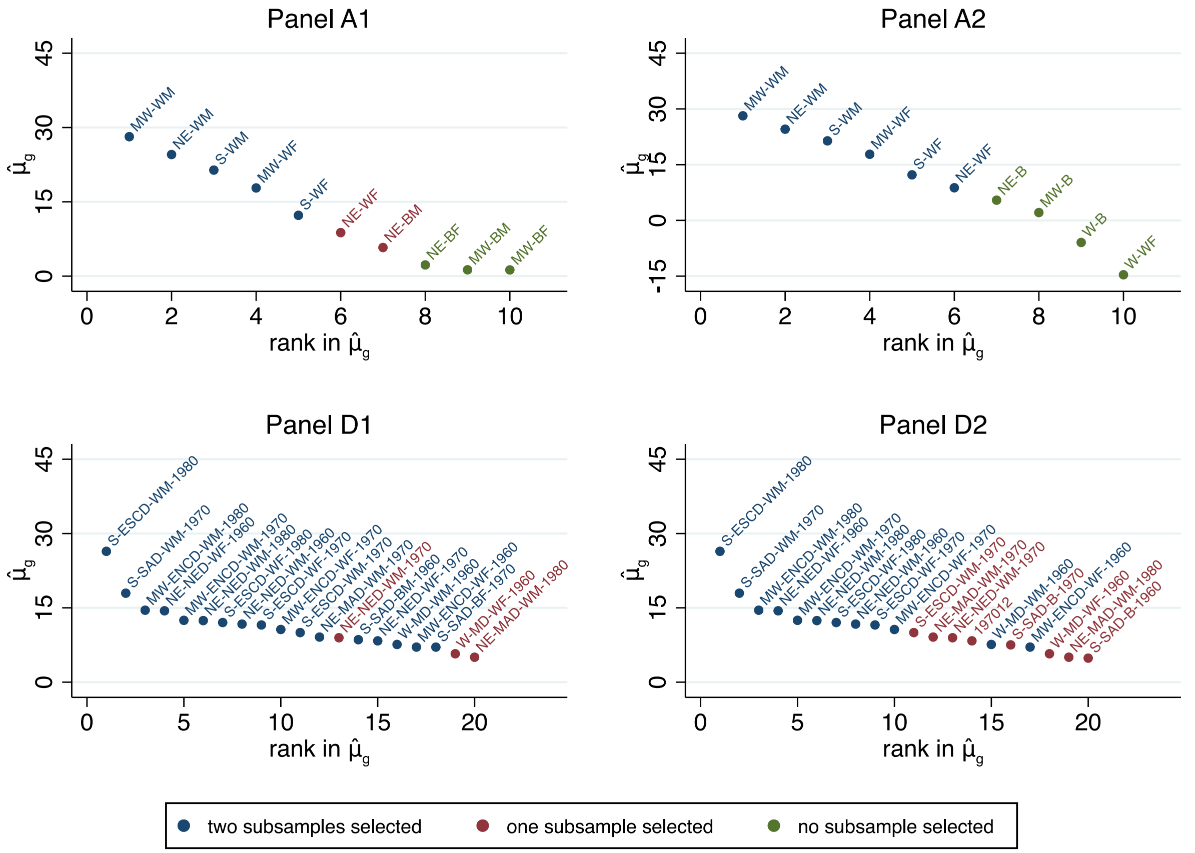

Fig. 3. Return to compulsory schooling: First-stage signal by groups. Note: Dataset is from Stephens and Yang (2014). The endogenous regressor is years of schooling. The instrument is the compulsory schooling year a birth cohort faces at age 14. All groupwise regressions also control state, survey year, census division by birth year, and census division by gender and race fixed effects. Regressions in the top two graphs also include a fourth-order polynomial in age. "NE", "MW", "S", and "W" in the labels of the top two figures stand for the Northeast, the Midwest, the South, and the West. "NE-NED", "NE-MAD", "MW-ENCD", "MW-WNCD", "S-SAD", "S-ESCD", "S-WSCD", "W-MD", and "W-PD" in the bottom two figures stand for the New England, the Middle Atlantic, the East North Central, the West North Central, the South Atlantic, the East South Central, the West South Central, the Mountain, and the Pacific Census divisions. "WM", "WF, "BM", and "BF" denote White males, White females, non-White males, and non-White females. "B" in panels A2 and B2 denotes the non-Whites as a robustness check. "1960", "1970", "1990", and "2000" denote Census survey year.

Let subscript c denote county, s denote state, and t denote year (when an election takes place). We consider the model,

$$\Delta Vote_{cst} = \beta \Delta Economy_{cst} + X_{cst}\theta_s + u_{cst}$$

$$\Delta Economy_{cst} = \sum_{s=1}^{S} \rho_s \Delta EnergySupply_t \times EnergyShare_{cs} + X_{cst}\gamma_s + v_{cst}, \tag{7}$$

where $\Delta Vote_{cst}$ is the change in voter turnout between two elections, $\Delta Economy_{cst}$ is local market activity measured by change in log per capita earning or change in log employment per adult, the instrument $\Delta EnergySupply_t \times EnergyShare_{cs}$ is the change in national employment level in oil and coal industries interacted with initial county-level employment share of the mining industry from the 1967 CBP, and X_{cst} is the list of exogenous regressors in Charles and Stephens (2013), which includes state-year fixed-effects as well as changes in time-varying county characteristics such as log total population, percentage of female adults, percentage of Black adults, percentage of other race and percentage of population aged 30s, 40s, 50s, 60s, and 70s and up.

Our definition of $EnergyShare_{cs}$ is different from that in Charles and Stephens (2013), which uses two dummy variables that indicate large and median employment share in oil or coal generated by 1974 CBP industry employment data. Because the sample spans from 1969 to 2000, using the 1974 CBP can potentially harm the validity of the exclusion restriction. On the other hand, the 1967 CBP employment measurement (also used in Charles and Stephens, 2013 for robustness checks) is not for the oil and coal industries, but for the entire mining industry. Hence, the 1967 CBP mining industry employment measure is expected to produce a weaker first stage than the 1974 CBP, which refers specifically to the oil and coal industries.

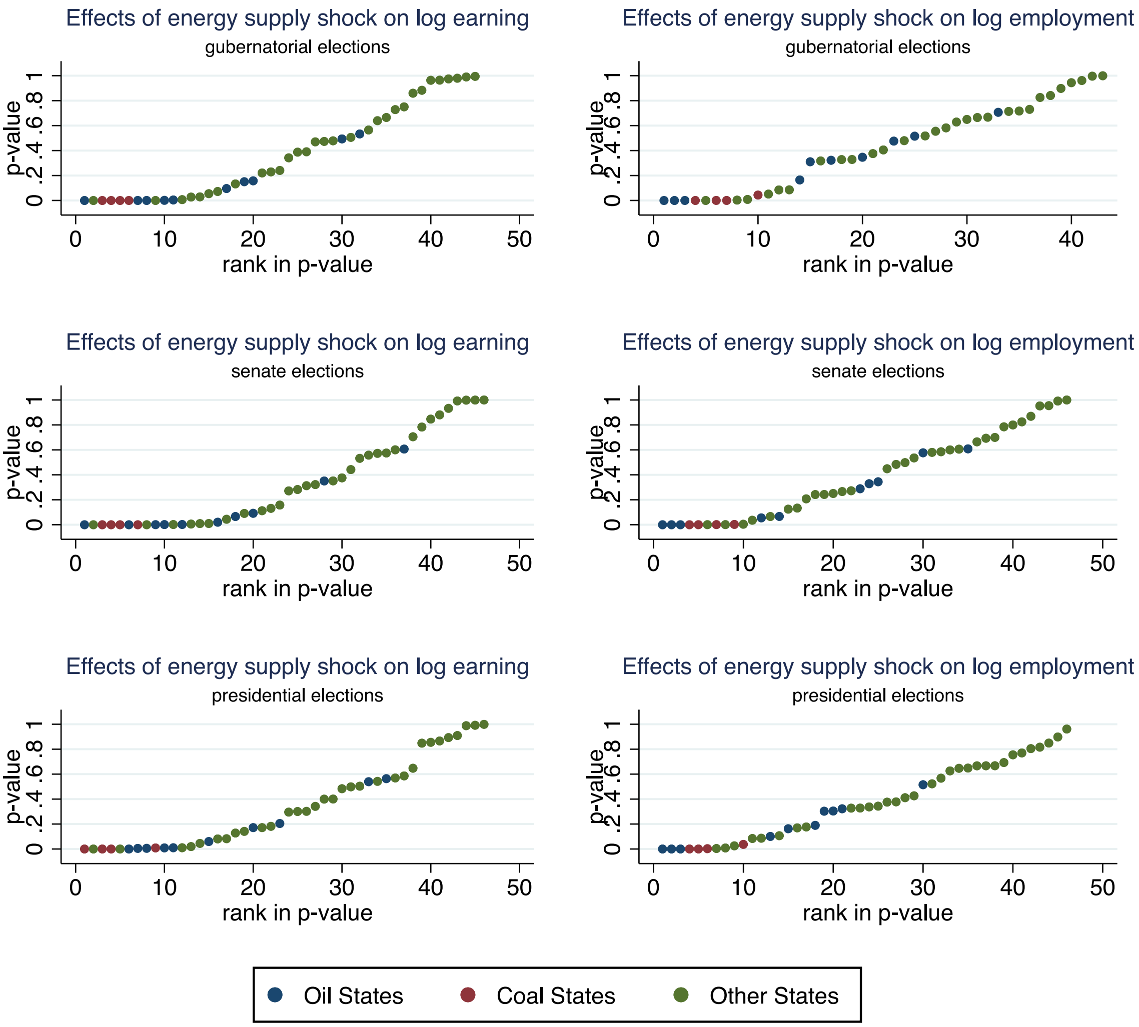

Fig. 4. Voter turnout: p-values of groupwise first-stage t-tests. Note: Dataset is from Charles and Stephens (2013). The endogenous regressor is change in log county-level per capita earning in the left column and change in log county-level employment per adult in the right column. The instrument is the change in national employment in oil/gas and coal interacted with the share of the mining industry in local employment in 1967. Other exogenous controls include county-year fixed-effects as well as changes in time-varying county characteristics such as log total population, percentage female adults, percentage Black adults, percentage "other" race and percentage population aged 30s, 40s, 50s, 60s, and 70s and up.

Indeed, the use of the 1967 employment measure instead of the 1974 measure to construct the instruments of the model in Eq. (7) generates non-significant pooled 2SLS results for the second-stage coefficient, β, even after restricting the sample to the fourteen oil/coal states defined in Charles and Stephens (2013). To preserve the exclusion restriction, in our analysis we employ the 1967 CBP industry employment data and define $EnergyShare_{cs}$ to be initial employment share in the mining industry. As we show below, β becomes significant when it is estimated with our adaptive procedure.

Graphs in Fig. 4 report p-values of groupwise one-sided first-stage t-tests. The graphs in the left column are for $Economy_{cst}$ equal to log per capita earning, while those in the right column are for $Economy_{cst}$ equal to log employment per adult. Row-wise, the graphs report results for gubernatorial, senate, and presidential elections, respectively, as indicated in the titles. All graphs show a mix between groups with strong and irrelevant instruments. Groups with strong first-stage identification give close to zero p-values, while groups with irrelevant instruments give near uniformly distributed p-values on the graphs.

Although all four coal states defined in Charles and Stephens (2013) seem to have a strong first stage, not all ten oil states have a strong first stage. Moreover, there are states other than the fourteen oil/coal states in Charles and Stephens (2013) that have a strong first-stage relationship between local labor market outcomes and energy supply shocks. Therefore, Fig. 4 provides ample motivation to apply our proposed methodology of selecting strong first-stage signals with the target of minimizing the asymptotic MSE.

Columns (1)–(8) of Table 6 report regression results from the same estimation methods as in Table 5. Column (9) reports results from pooled 2SLS using the fourteen pre-determined oil and coal states defined in Charles and Stephens

Table 6
Effects of local economic performance on voter turnout: Estimation results.

Full-sample				Select-and-interact				
OLS	2SLS-P	2SLS-INT	LIML-INT	2SLS-SSL	2SLS-ADPT			2SLS-CS
					(κ^*)	($2\kappa^*$)	($\kappa^*/2$)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panels A1–C1: log per-capita earning								
Panel A1: Gubernatorial elections								
−0.001	−0.024	-0.038**	-0.044***	-0.046*	-0.056**	-0.057**	-0.050*	−0.020
(0.002)	(0.020)	(0.015)	(0.016)	(0.023)	(0.025)	(0.028)	(0.026)	(0.021)
Panel B1: Senate elections								
0.007***	−0.017	-0.026**	-0.032**	-0.054***	-0.039**	-0.038**	-0.038**	−0.029
(0.002)	(0.017)	(0.013)	(0.014)	(0.020)	(0.016)	(0.017)	(0.016)	(0.018)
Panel C1: Presidential elections								
0.002	−0.018	−0.007	−0.009	−0.025	−0.108	−0.106	−0.104	−0.003
(0.001)	(0.021)	(0.015)	(0.016)	(0.030)	(0.067)	(0.067)	(0.067)	(0.024)
Panels A2–C2: log employment per adult								
Panel A2: Gubernatorial elections								
0.008**	−0.065	-0.105***	-0.117***	−0.189	-0.152***	-0.148***	-0.161***	−0.045
(0.004)	(0.054)	(0.030)	(0.032)	(0.118)	(0.045)	(0.049)	(0.043)	(0.048)
Panel B2: Senate elections								
0.020***	−0.049	-0.066**	-0.085**	-0.111***	-0.118***	-0.123***	-0.140***	−0.070
(0.004)	(0.049)	(0.029)	(0.032)	(0.038)	(0.039)	(0.039)	(0.040)	(0.043)
Panel C2: Presidential elections								
0.023***	−0.050	−0.023	−0.036	−0.186	-0.166*	-0.166*	-0.194**	−0.007
(0.003)	(0.060)	(0.036)	(0.041)	(0.142)	(0.092)	(0.092)	(0.098)	(0.055)

Note: Dataset is from Charles and Stephens (2013). The endogenous regressor is change in log county-level per capita earning in Panel A and change in log county-level employment per adult in Panel B. The instrument is the change in national employment in oil/gas and coal interacted with the share of the mining industry in local employment in 1967. Other exogenous controls include county-year fixed-effects as well as changes in time-varying county characteristics such as log total population, percentage female adults, percentage Black adults, percentage "other" race and percentage population aged 30s, 40s, 50s, 60s, and 70s and up.

(2013). Full-sample pooled 2SLS and pooled 2SLS with data from oil and coal states produces statistically insignificant results across all six rows. This could be caused by the use of the 1967 crude measure of local employment to construct the instrument. In contrast, all regressions in columns (3)–(8) utilize the heterogeneity in first-stage model. Results in these columns are generally negative and statistically significant for the gubernatorial and senate elections.

All estimators reported in columns (3)–(8) are first-order equivalent under the assumptions in Section 3. The fully-interacted 2SLS (2SLS-INT) has the smallest asymptotic variance, but could suffer from non-trivial many-IV bias. The proposed adaptive method (2SLS-ADPT) and the repeated split-sample interacted 2SLS lasso estimator (2SLS-SSL) have larger higher order asymptotic variances compared to the fully-interacted 2SLS, but also enjoy better rates in higher order asymptotic bias. The results in Table 6 are consistent with the formal properties of the estimators. The 2SLS-INT in column (3) has smaller standard errors than the split-sample adaptive estimators in columns (5)–(8). However, point estimates in column (3) fall between the OLS estimates in column (1) and the split-sample selective estimates in columns (5)–(8), providing evidence of many-IV bias in the direction of OLS. Similar to the first empirical application, the proposed adaptive procedure gives results that are robust to perturbations in the tuning parameter sequence $\kappa_{G,N}$. The results for 2SLS-SSL are similar to those of 2SLS-ADPT procedure in panels A1, B1, and B2, but they appear to be less precise than the 2SLS-ADPT results in panels A2 and C2.

6. Conclusion

In this article, we study a linear simultaneous equation model with a scalar endogenous regressor, an external instrument, and a heterogeneous first-stage relationship between the endogenous regressor and the instrument that varies across groups. This is a natural set-up in many empirical applications in economics. Under first-stage heterogeneity, pooled 2SLS is inefficient. 2SLS using the interactions between the external instrument and the full set of group dummies as IV suffers from many-IV bias. We show that sample selection based on the first-stage correlation coupled with pooled 2SLS, a strategy seen in some applied studies, yields invalid inference. Sample selection followed by interacted 2SLS preserves first-order efficiency but may still have substantial higher-order asymptotic bias. Following earlier work by Donald and Newey (2001) and others, we propose a data-driven procedure for the selection of groups in the sample. Our procedure is designed to minimize the high-order MSE expansion of the second-stage estimator.

Although our set-up assumes a homogeneous second stage to facilitate the asymptotic MSE comparison, our proposed estimator has a weighted average causal effect type of interpretation when the second stage is heterogeneous across groups. We show that, for the weights to be positive and for the estimator to be invariant to groupwise rescalings of the instrument, it is crucial to interact the external instrument *as well as* all exogenous controls with the full set of group dummies.

Our adaptive procedure is akin to a version of the split-sample IV lasso of Belloni et al. (2012) applied to the case when the first-stage regressors are interactions between an instrument and group indicators and interactions between all exogenous controls and group indicators. Our allowance for a non-zero proportion of weak instruments is similar to their approximate sparsity condition. When the proportion of weak instruments goes to zero, our adaptive estimator is asymptotically equivalent to the split-sample IV lasso estimator, because both methods consistently select all groups with strong instruments in the first-stage. When the proportion of weak instruments does not go to zero, both estimators are consistent and asymptotic normal. We are not aware of higher-order asymptotic MSE analyses for the IV lasso estimator or the split-sample IV lasso estimator. In simulations, we compare the performance of our proposed adaptive estimator to the pooled 2SLS and fully-interacted 2SLS estimators popular in the applied literature, as well as the split-sample IV lasso, LIML, UJIVE, and IJIVE estimators with a comparable fully-interacted specification. We find that our proposed adaptive estimator performs significantly better than the pooled 2SLS and fully-interacted 2SLS estimators when the proportion of groups with strong first-stage signal is small. In addition, we find that the proposed adaptive estimator performs as well as, and in some cases better than, the other competing estimators.

We apply our proposed methods to study *(i)* the return to compulsory schooling, and *(ii)* the effect of local labor market conditions on voter turnout, following Stephens and Yang (2014) and Charles and Stephens (2013), respectively. We show that taking into account first-stage heterogeneity improves statistical precision in both applications. In contrast to the results in Stephens and Yang (2014), our proposed procedure produces statistically significant estimates of 3–4 percent for the effect of an additional year of schooling on wages, even after controlling for demographic region by birth cohort fixed effects. In the second application, efficiency gains obtained through our group selection procedure allow us to replicate the main results of Charles and Stephens (2013) using an alternative sample with a weaker instrument, but a more plausible exclusion restriction.

Acknowledgment

Jiaying Gu acknowledges financial support from Social Sciences and Humanities Research Council of Canada.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2023.02.005.

References

Abadie, A., 2003. Semiparametric instrumental variable estimation of treatment response models. J. Econometrics 113, 231–263.

Acemoglu, D., Finkelstein, A., Notowidigdo, M.J., 2013. Income and health spending: Evidence from oil price shocks. Rev. Econ. Stat. 95 (4), 1079–1095.

Acemoglu, D., Johnson, S., Robinson, J., Yared, P., 2008. Income and democracy. Amer. Econ. Rev. 98 (3), 808–842.

Ackerberg, D.A., Devereux, P.J., 2009. Improved JIVE estimators for overidentified linear models with and without heteroskedasticity. Rev. Econ. Stat. 91 (2), 351–362.

Altmejd, A., Barrios-Fernández, A., Drlje, M., Goodman, J., Hurwitz, M., Kovac, D., Mulhern, C., Neilson, C., Smith, J., 2021. O brother, where start thou? Sibling spillovers on college and major choice in four countries. Q. J. Econ. 136 (3), 1831–1886.

Andrews, I., Armstrong, T.B., 2017. Unbiased instrumental variables estimation under known first-stage sign. Quant. Econ. 8 (2), 479–503.

Angrist, J., Imbens, G., 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. J. Amer. Statist. Assoc. 90, 431–442.

Angrist, J.D., Imbens, G.W., Krueger, A.B., 1999. Jackknife instrumental variables estimation. J. Appl. Econometrics 14 (1), 57–67.

Angrist, J.D., Pischke, J.S., 2009. Mostly Harmless Econometrics. Princeton University Press.

Bekker, P.A., 1994. Alternative approximations to the distributions of instrumental variable estimators. Econometrica 62 (3), 657–681.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 (6), 2369–2429.

Berkowitz, D., Caner, M., Fang, Y., 2008. Are nearly exogenous instruments reliable? Econom. Lett. 101 (1), 20–23.

Black, D., Daniel, K., Sanders, S., 2002. The impact of economic conditions on participation in disability programs: Evidence from the coal boom and bust. Amer. Econ. Rev. 92 (1), 27–50.

Bound, J., Jaeger, D.A., Baker, R.M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J. Amer. Statist. Assoc. 90, 443–450.

Card, D., Devicienti, F., Maida, A., 2014. Rent-sharing, holdup, and wages: Evidence from matched panel data. Rev. Econom. Stud. 81 (1), 84–111.

Cervellati, M., Jung, F., Sunde, U., Vischer, T., 2014. Income and democracy: Comment. Amer. Econ. Rev. 104 (2), 707–719.

Charles, K.K., Li, Y., Stephens, Jr., M., 2018. Disability benefit take-up and local labor market conditions. Rev. Econ. Stat. 100 (3), 416–423.

Charles, K., Stephens, Jr., M., 2013. Employment, wages, and voter turnout. Am. Econ. J. 5 (4), 111–143.

Cheng, X., Liao, Z., Shi, R., 2019. On uniform asymptotic risk of averaging GMM estimators. Quant. Econ. 10 (3), 931–979.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., 2017. Double/debiased/neyman machine learning of treatment effects. Amer. Econ. Rev. 107 (5), 261–265.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21 (1), 1–68.

Chernozhukov, V., Hansen, C., Spindler, M., 2016. Hdm: High-dimensional metrics. arXiv preprint arXiv:1608.00354.

Coussens, S., Spiess, J., 2021. Improving inference from simple instruments through compliance estimation. arXiv preprint arXiv:2108.03726.

Currie, J., Moretti, E., 2003. Mother's education and the intergenerational transmission of human capital: Evidence from college openings. Q. J. Econ. 118 (4), 1495–1532.

Deryugina, T., Heutel, G., Miller, N.H., Molitor, D., Reif, J., 2019. The mortality and medical costs of air pollution: Evidence from changes in wind direction. Amer. Econ. Rev. 109 (12), 4178–4219.

Dix-Carneiro, R., Kovak, B.K., 2017. Trade liberalization and regional dynamics. Amer. Econ. Rev. 107 (10), 2908–2946.

Donald, S.G., Newey, W.K., 2001. Choosing the number of instruments. Econometrica 69 (5), 1161–1191.

Fredriksson, P., Ockert, B., Oosterbeek., H., 2013. Long-term effects of class size. Q. J. Econ. 128 (1), 249–285.

Guggenberger, P., 2012. On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. Econom. Theory 28 (2), 387–421.

Hahn, J., Hausman, J., Kuersteiner, G., 2004. Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. Econom. J. 7 (1), 272–306.

Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.

Jackson, C.K., Johnson, R.C., Persico, C., 2016. The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. Q. J. Econ. 131 (1), 157–218.

Johnson, M.S., 2020. Regulation by Shaming: Deterrence effects of PublicizingViolations of workplace safety and health laws. Am. Econ. Rev. 110 (6), 1866–1904.

Kolesár, M., 2013. Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity. Unpublished Working Paper.

Kolesár, M., 2018. Minimum distance approach to inference with many instruments. J. Econometrics 204 (1), 86–100.

Kuersteiner, G., Okui, R., 2010. Constructing optimal instruments by first-stage prediction averaging. Econometrica 78 (2), 697–718.

Leeb, H., Pötscher, B., 2005. Model selection and inference: Facts and fiction. Econom. Theory 21 (1), 21–59.

Lleras-Muney, A., 2002. Were state laws on compulsory education effective? An analysis from 1915 to 1939. J. Law Econ. 45 (2), 401–435.

Lleras-Muney, A., 2005. The relationship between education and adult mortality in the united states. Rev. Econom. Stud. 72, 189–221.

Nagar, A., 1959. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. Econometrica 27 (4), 575–595.

Okui, R., 2009. The optimal choice of moments in dynamic panel data models. J. Econometrics 151 (1), 1–16.

Oreopoulos, P., 2006. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. Amer. Econ. Rev. 96 (1), 152–175.

Pascali, L., 2017. The wind of change: Maritime technology, trade, and economic development. Amer. Econ. Rev. 107 (9), 2821–2854.

Phillips, G.D.A., Hale, C., 1997. The bias of instrumental variable estimators of simultaneous equation systems. Internat. Econom. Rev. 18 (1), 219–228.

Sølvsten, M., 2020. Robust estimation with many instruments. J. Econometrics 214 (2), 495–512.

Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586.

Stephens, M., Yang, D.-Y., 2014. Compulsory education and the benefits of schooling. Amer. Econ. Rev. 104 (6), 1777–1792.

Stock, J., Yogo, M., 2005. Asymptotic distributions of instrumental variables statistics with many instruments. In: Andrews, D., Stock, J. (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. Cambridge University Press, pp. 109–120.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Amer. Statist. Assoc. 113, 1228–1242.