Instrumental Variable Estimation with First-Stage Heterogeneity

Alberto Abadie
MIT

Jiaying Gu
University of Toronto

Shu Shen
University of California, Davis

July 2022

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.

Alberto Abadie, Department of Economics, MIT, abadie@mit.edu. Shu Shen, Department of Economics, UC Davis, shushen@ucdavis.edu. Jiaying Gu, Department of Economics, University of Toronto, jiaying.gu@utoronto.ca. 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!

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 constructed based on policy changes as natural experiments. In this type of setting, variations in details of the policy or its level of enforcement often result in natural first-stage heterogeneity across geographic regions (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, Johnson, and Persico, 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. In the literature on the return to compulsory schooling, for example, 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 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. This study 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

[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) justifies 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".

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. Cervellati, Jung, Sunde, and Vischer (2014) argues that the instrument used in an influential article by Acemoglu, Johnson, Robinson, and Yared (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 tends to generate overly large biases of second-stage instrumental variable estimators, and overly large second-stage t-statistics under the null in significance tests. Using different data samples for sample selection and 2SLS estimation (e.g., the U.S. analysis in Altmejd, Barrios-Fernández, Drlje, Goodman, Hurwitz, Kovac, Mulhern, Neilson, and Smith, 2021) is a much better practice, but the method is generally inefficient, as we discuss below.

An alternative empirical approach adopted in empirical research exploits 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, Heutel, Miller, Molitor, and Reif (2019) interact wind direction with pollution-monitor geo-cluster indicators to instrument for air pollution. Jackson, Johnson, and Persico (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. Similarly, Pascali (2017) allows for time-varying first-stage coefficients when utilizing 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, the method may suffer from misleading inference due to the "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 a 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 of groupwise IV strength is equivalent to carrying out 2SLS interacting the instrument with the full set of group dummy variables. Our proposed estimator improves upon the fully interacted estimator by employing multiple testing of instrument strength in each group and using the asymptotic MSE of the second-stage estimator as the criteria to form the decision rule of the first-stage tests. We propose a procedure where the cut-off value for first-stage testing is adaptively chosen to minimize the asymptotic MSE of the second-stage estimator. Sample splitting following, for example, Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017); Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018); Wager and Athey (2018) is adopted in the proposed adaptive procedure to separate first-stage testing from second-stage estimation to reduce asymptotic bias.

This article 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.

Our set-up assumes a homogeneous second-stage to facilitate the comparison of different estimation approaches under the asymptotic MSE framework. When the second stage is heterogeneous, our proposed estimator has an interpretation as a weighted average causal

effect. Competing efficient estimators with the same order in the higher order term of the MSE formula, such as the limited information likelihood estimator (LIML), the Jackknife IV estimator (JIVE), or the bias-corrected 2SLS estimator (B2SLS) do not have such interpretation. In addition, we show that when the IV is allowed to have heterogeneous variation across groups, our proposed estimator has an additional appealing property of being invariant to groupwise rescaling of the IV, compared to many other competing estimators.

Our proposed estimation procedure focuses on higher-order asymptotic expansion of the MSE to choose instruments, related to earlier work of Donald and Newey (2001), Okui (2009), Cheng, Liao, and Shi (2019). The key difference between the MSE expansion literature and the instrument selection strategy in the machine learning literature (e.g., Belloni, Chen, Chernozhukov, and Hansen, 2012; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018) is that the IV selection criteria of the former is based on the asymptotic MSE of the second-stage estimator while 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 close to a modified version of the split-sample IV lasso method with groupwise interacted instruments. Our paper is also related to contemporary work by Coussens and Spiess (2021), who study the evaluation of a binary endogeneity treatment when a randomized IV is available, and proposes to reweigh 2SLS by individual compliance propensity estimated through cross-validated flexible nonparametric tools, such as causal forests. Like Coussens and Spiess (2021), we study the possibility of utilizing first-stage heterogeneity to improve the precision of second-stage IV estimation. When the endogenous regressor in our model is binary, the first-stage heterogeneity we consider reduces to the compliance propensity discussed in Coussens and Spiess (2021). In addition, unlike Coussens and Spiess (2021) and motivated by empirical practice, we explicitly consider the potential presence of subsamples that might lack first-stage identification, and propose to discard such groups. Our procedure selects the discarded groups adaptively to minimize high-order terms in the asymptotic MSE formula of the second-stage estimator.

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 on the return to schooling. Once they control for regional cohort fixed effects, Stephens and Yang (2014) obtain IV estimates of return to schooling that are not statistically significant. Using similar model specifications, we find that while 2SLS over the entire sample produces statistically insignificant results, after allowing for 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. 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 data set contains detailed industry-level information at the county level, instruments based on these data may not provide completely exogenous variation for the 1964-2000 estimation window in Charles and Stephens (2013). As a robustness check, Charles and Stephens (2013) use the 1969 CBP data. The 1969 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 1969 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 1969 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 1969 CBP data produces 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 non-trivial, weak, or zero for different population subgroups. We discuss asymptotic properties of the existing methods and particularly 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 covariate, 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 the exogenous regressors. In empirical research, groups could be determined by observables like geographic regions, ethnic groups, etc. To facilitate the comparison of different estimators in an asymptotic MSE framework, we initially assume that β is a constant. 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 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* $\underline{c}\frac{N}{G} \leq n_g \leq \bar{c}\frac{N}{G}$ *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* $k_g = E[\eta_{ig}]$ *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* $g = 1, \ldots, G$ *and* $E[(\eta_{ig} - k_g)^2] \leq \bar{\Delta}_\eta < \infty$ *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* $G, N \to \infty$, $G_{+,s}/G \to b > 0$.

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. In our two empirical applications, where compulsory schooling laws instrument for years of education and energy supply shocks instrument for local economic conditions, 1.2 is a natural assumption. When the first-stage coefficient is of order $1/\sqrt{N/G}$, we say that the instrument is weak.

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 important in some empirical applications including the two examples we study later in the paper.

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, Liao, and Shi (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 $g = 1, 2, ..., G$. 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 $\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_G$, D the $N \times G$ block diagonal matrix of $Z_1, ..., Z_G$, D_X the $N \times G$ block diagonal matrix of $X_1, ..., X_G$, and D_ℓ the $N \times G$ block diagonal matrix of $\ell_1, ..., \ell_G$. 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,

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

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 can be important to the interpretation of 2SLS estimators, as we will discuss in the next section.

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

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

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

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

Both estimators are consistent. The fully interacted estimator is more efficient since $s_{int} \leq s_p$ 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 $Z_{inf} = (\rho_1 Z_1', \ldots, \rho_G Z_G')'$ 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 the "many IV bias" as is discussed in Bekker (1994), Bound, Jaeger, and Baker (1995), Staiger and Stock (1997), and Stock and Yogo (2005), among many others.

2.2.2. Interpretation Under Second-Stage Effect Heterogeneity

As we discuss in the introduction, IV methods such as LIML, JIVE, or B2SLS have been proposed in the econometrics literature to preserve estimation efficiency under homoskedas-

ticity while addressing the "many IV bias" problem of 2SLS. Nonetheless, 2SLS is still the most popular method in empirical research, perhaps because 2SLS has a weighted average interpretation when the second-stage causal effect is not constant. 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 (see, e.g., Imbens and Angrist., 1994, and Abadie, 2003). This subsection studies the interpretation of pooled and fully-interacted 2SLS estimators when a heterogeneous second-stage causal parameter is added to our model in (1) with first-stage heterogeneity. 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 $|\beta_g| \leq \bar{\beta} < \infty$ for all $g = 1, ..., G$. 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 pooled and fully interacted estimators differ in that 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 factor of 10, the variance of the instrument in group g increases by a factor of 100, but the first-stage slope coefficient ρ_g is divided by a factor of 10 only. This seemingly harmless transformation changes the interpretation of the pooled estimator but not 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 group-specific intercepts, 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 $a_g = E[\tilde{Z}_{ig}^2]$ and $b_g = E[\tilde{Z}_{ig}]$. In the appendix, we show $\hat{\beta}_{pool2} = (\sum_{g=1}^{G} \rho_g p_g (a_g - b_g \sum_{s=1}^{G} b_s p_s))^{-1} (\sum_{g=1}^{G} \beta_g p_g (\gamma (b_g - \sum_{s=1}^{G} b_s p_s) + \rho_g (a_g - b_g \sum_{s=1}^{G} b_s p_s))) + o_p(1)$, 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.

2.2.3. Naive First-Stage Selection

A selective IV regression approach is sometimes used by applied researchers aiming to obtain a strong first stage. If subsamples are selected for IV regression based on economic intuition before seeing the data (e.g., Fredriksson, Ockert, and Oosterbeek., 2013; Card, Devicienti, and Maida, 2014), the selective IV approach may be legitimate. However, when the selection is based on first-stage regression results, the process turns out to invalidate 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 $H_{a,g} : \rho_g > 0$, $g = 1, ..., G$. 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 the estimator as the select-and-pool estimator.

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

Theorem 1. *Suppose Assumption 1 holds, $\sigma_{uv} \neq 0$, and $G_{+,s}/G \to b \leq \bar{b} < 1$ as $G, N \to \infty$. 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 exclusion restriction is violated for any finite sample if the select-and-pool method is employed. This is because the selection is based on the value of the first-stage t-statistic in each group, and a subgroup is more likely to be selected when there is a large positive correlation between the instrument and the first-stage error term. Since 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 size, 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, Caner, and Fang (2008), Guggenberger (2012) among others). In this regime of local violation, classical inference starts to fail for many IV estimators. For instance Berkowitz, Caner, and Fang (2008) shows that 2SLS has a limiting distribution

[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).

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 rate ranges from 7 to 10 percent. 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 is illustrated in Table A1 in the Appendix, which reports standard deviations for the different estimators. The fully interacted estimator suffers from the "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). With our DGP, the "many IV bias" is a finite sample problem. So 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.

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. On the other hand,

Table 1: Size-Distortion 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 the rejection proportion 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 $g = 1, ..., G_{+,s}$ and $\rho_g = 0$ for $g > G_{+,s}$.

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 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 this requirements and is amenable to classic asymptotic inference.

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 fist-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 $\hat{\beta}_{int}(\delta)$. 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 then 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 samples of equal proportions within each group. We use superscripts a and b to refer to the observations in the two sample splits. Let $\hat{\rho}_g^a = \left((Z_g^a)'Z_g^a)\right)^{-1}(Z_g^a)'W_g^a$, $\hat{\mu}_g^a = \left((Z_g^a)'Z_g^a)\right)^{-1/2}(Z_g^a)'W_g^a$ and define similar terms for subsample b. Let

$$\begin{aligned}
\hat{\beta}^a(\delta) &= \left(\sum_g \hat{\rho}_g^b (Z_g^a)'W_g^a 1(\hat{\mu}_g^b \geq \delta)\right)^{-1} \sum_g \hat{\rho}_g^b (Z_g^a)'Y_g^a 1(\hat{\mu}_g^b \geq \delta), \\
\hat{\beta}^b(\delta) &= \left(\sum_g \hat{\rho}_g^a (Z_g^b)'W_g^b 1(\hat{\mu}_g^a \geq \delta)\right)^{-1} \sum_g \hat{\rho}_g^a (Z_g^b)'Y_g^b 1(\hat{\mu}_g^a \geq \delta), \text{ and} \\
\hat{\beta}_{sssel,int}(\delta) &= \left(\hat{\beta}^a(\delta) + \hat{\beta}^b(\delta)\right)/2. \qquad (5)
\end{aligned}$$

$\hat{\beta}^a(\delta)$ 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. 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, as we show below.

The following Assumption gives a range condition for δ.

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

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 Assumption 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.[3] 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 a function of δ.

3.2. Characterization of Asymptotic Mean Squared Errors

To approximate the MSEs of the select-and-interacted 2SLS estimators as a function of the value δ, we apply the higher-order asymptotic expansion techniques in Nagar (1959), Donald and Newey (2001), Okui (2009), Cheng, Liao, and Shi (2019), and others. To keep the calculation tractable, we assume in this section that the error terms (u, v) follow a joint normal distribution. Let $\Phi(.)$ and $\phi(.)$ be the cumulative distribution function and the probability density function of the standard normal distribution function, respectively.

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

[3]Efficiency follows by comparison between s_{int} and $s_{sel,int}$, and $\sum_{g \in \mathcal{G}^c_{+,s}} \rho_g^2 k_g p_g = O_p(G/N)$.

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*

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

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* $G/N \to 0$ *as* $G, N \to \infty$*, the asymptotic MSE of* $\hat{\beta}_{sssel,int}$ *can be decomposed as*

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

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

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

For the full-sample select-and-interact estimator $\hat{\beta}_{sel,int}$, the first-order term in its asymptotic MSE decomposition is the variance term σ_u^2/H, which is expected given the asymptotic variance formula in Lemma 1. The higher-order terms of the estimator may come from three different sources, the "many IV bias", the bias introduced by first-stage selection, and the efficiency loss from falsely excluding groups with relevant instruments. Under the δ range 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 the first-stage selection were orthogonal to the second-stage estimation, the bias term would be $\sigma_{uv}^2 \Big(\sum_g \left(1 - \Phi\left(\frac{\delta - \mu_g}{\sigma_v} \right) \right) \Big)^2 / N$.

The additional terms in $S_{sel,int}(\delta)$ hence represent extra asymptotic bias introduced from first-stage selection.

The repeated split-sample estimator $\hat{\beta}_{sssel,int}$ has the same first-order 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 higher order MSE terms are from two different sources. $A_{sssel,int}(\delta)$ represents the higher order bias of the split-sample estimators. $B_{sssel,int}(\delta)$ represents higher-order efficiency loss from excluding groups with weak instruments. $C_{sssel,int}(\delta)$ is a higher-order bias 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 $L(\delta) = H^2 S_{sssel,int}(\delta)$.*

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 vaninsh, 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 estimator as discussed in Leeb and Pötscher (2005). As a result, we do not expect characterizing the minimum asymptotic MSE level of $\hat{\beta}_{sssel,int}$ in the not well-separated case would lead to meaningful adaptive procedure for choosing the optimal thresholding value δ given data.

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 op-

timal 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

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

Let the adaptive estimator be $\hat{\beta}_{adpt} \equiv \hat{\beta}_{sssel,int}(\hat{\delta})$ with $\hat{\delta}$ defined in Theorem 3. 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. The leading higher order terms still converges to L^*, but L^* may not necessarily be the smallest among all split-sample select-and-interact estimators defined in equation (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 has to be 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 that will include all strong groups but discard all other groups in the

limit. 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 $\kappa_{G,N}$ choices using $2\kappa^*_{G,N}$ and $\kappa^*_{G,N}/2$ and find the empirical results stable to such perturbations.

Theorem 3 implies that, under second-stage parameter heterogeneity (like in section 2.2.2), the adaptive estimator $\hat{\beta}_{adpt}$ satisfies

$$\hat{\beta}_{adpt} = \sum_{g \in \mathcal{G}_{+,s}} \frac{\rho_g^2 V_g p_g}{\sum_{g \in \mathcal{G}_{+,s}} \rho_g^2 V_g p_g} \beta_g + o_p(1). \tag{6}$$

Under correct selection of the strong first-stage groups, the adaptive estimator is equivalent to the oracle estimator that employs the identity of the groups with a strong first-stage identification, that is, $\hat{\beta}_{adpt} = \hat{\beta}_{oracle}$, where

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

It is clear that $\hat{\beta}_{oracle}$ has probability limit $\sum_{g \in \mathcal{G}_{+,s}} \frac{\rho_g^2 V_g p_g}{\sum_{g \in \mathcal{G}_{+,s}} \rho_g^2 V_g p_g} \beta_g$, following the same arguments as in Section 2.2.2. 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 adaptive estimator in (6) is valid for the adaptive estimator.

Our proposed adaptive procedure is akin to a version of the split-sample lasso selection estimator of Belloni, Chen, Chernozhukov, and Hansen, 2012. In simulations, we find that our proposed adaptive estimator behaves comparably and in some DGPs better than split-sample lasso. 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.

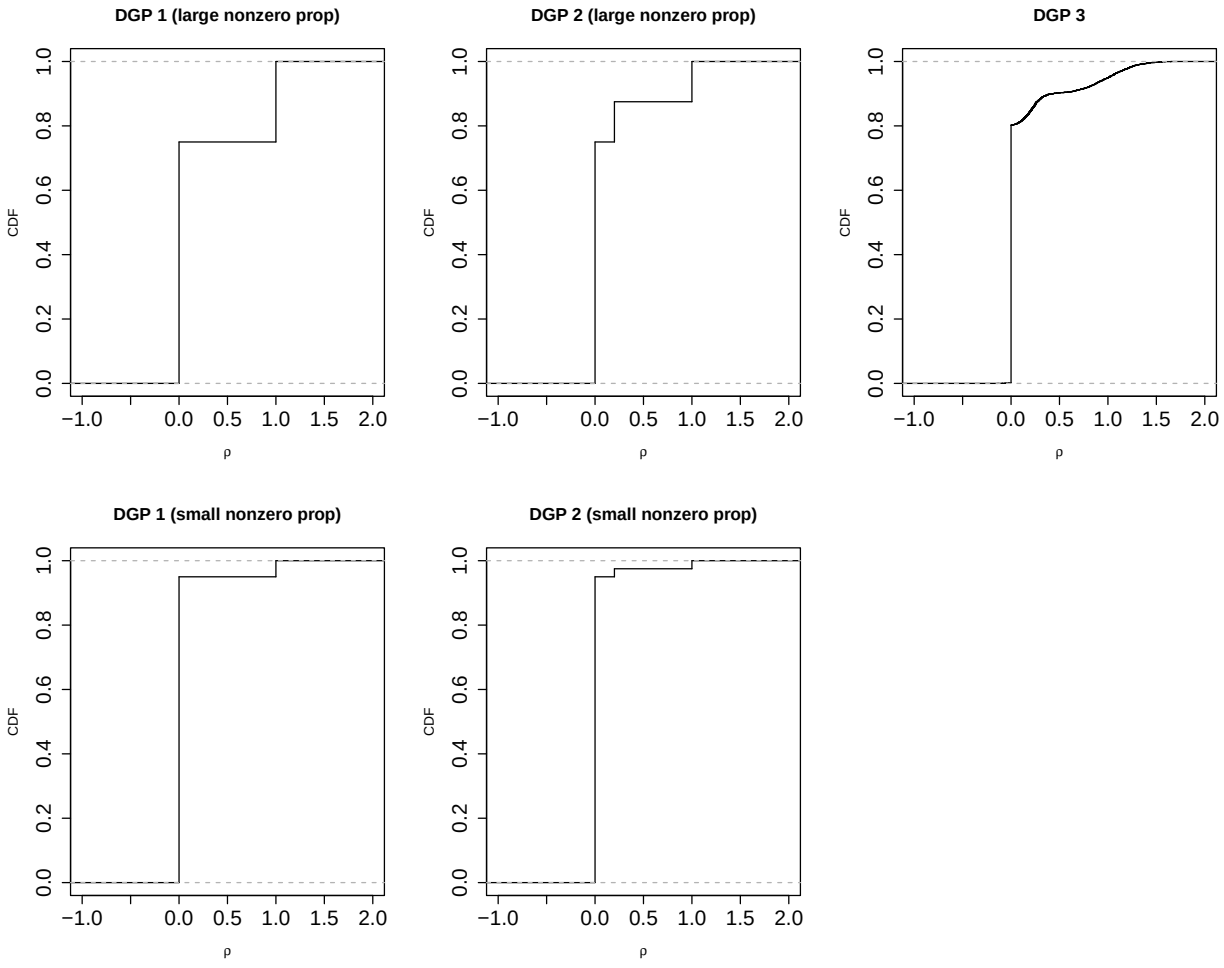
4. Monte Carlo Simulations

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 $\rho_{u,v}$. Endogenous variables Y_{ig} and W_{ig} are generated following the simultaneous equation model in (1)

with $\beta = 0$, and $\theta = \gamma = 1$. The parameter ρ_g controls the relevance of instrument Z in group g and varies across DGPs.

Figure 1 summarizes the distribution of ρ_g for the three DGPs. 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 Figure 1 plot the cumulative distribution functions (CDFs) of DGP 1 with p_s equal to 0.25 and 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 Figure 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. Eighty percent of the groups in DGP 3 have irrelevant instruments. Among remaining twenty 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.

Figure 1: Distribution of ρ_g for three DGPs



In our simulations, we study the performance of the following estimators: (1) $\hat{\beta}_{pool}$ (2SLS-P) the conventional 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 which interacts the instrument only with groups that have non-zero first-stage correlation, denoted as 2SLS-INF, (5) the limited information maximum likelihood estimator (LIML-INT) with fully interacted instrument with group dummies, (6) The jack-knife estimator (JIVE), which was first proposed by Angrist, Imbens, and Krueger (1999) to alleviate finite-sample bias of 2SLS, (7) a

split-sample 2SLS estimator that uses lasso for first-stage selection (2SLS-SSL), and (8) our 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 estimators $\hat{\beta}_{sel,int}(\infty)$ and $\hat{\beta}_{sssel,int}(\infty)$ defined in Section 3.1, respectively.

Tables 2-4 report empirical MSE and MAE (median absolute error) across 500 simulations, as well as rejection proportions for the second-stage t-test for the three DGPs under two different error distributions: normal and chi-squared with 3 degress of freedom (χ^2_3). For all DGPs, the pooled two stage least square estimator (2SLS-P) has very poor MSE and MAE performance. This is mainly driven by variance inflation: the large number of groups with no first stage makes the pooled 2SLS estimator inefficient. 2SLS-INT is first-order efficient. It behaves well when the number if 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 of many IV bias, 2SLS-INT can have much higher MSE than 2SLS-INF when G is large and the proportion of nonzero group is small.[4] In such cases, 2SLS-SSINT help to reduce the asymptotic bias, but is not as optimal in terms of asymptotic MSE as 2SLS-INF. 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 and JIVE relative to the infeasible estimator 2SLS-INF.

For DGP1 with normal errors, the proposed adaptive estimator is almost as good as the corresponding infeasible oracle estimator 2SLS-INF. All the other three competitors: 2SLS-SSL, LIML-INT and JIVE behave similarly. LIML-INT is a very competitive estimator under normal error which is not surprising given its optimality as discussed in Kolesár (2018). However, as discussed in Kolesár (2013), when treatment effects vary across groups LIML-INT does not have a clear causal interpretations, while JIVE also needs additional adjustment in this setting and has the highest computation burden.

For the same DGP but with χ^2_3 error distribution and a small proportion of non-zero

[4]In some settings, especially when the proportion of strong groups is large, 2SLS-INT can perform similar to 2SLS-INF. For example, under DGP1 with $p_s = 0.5$ (not reported), 2SLS-INT has similar MSE and MAE 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.

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 and JIVE by a substantial margin. When $p_s = p_w = 0.125$, they perform similarly. We also note there that the LIML-INT has noticeable size inflation whenever the error term is not normal. This phenomenon has been previously documented by Hahn, Hausman, and Kuersteiner (2004) and Sølvsten (2020), among others. Results for DGP2, which adds a small proportion of weak first-stage groups, are reported in Table 3. Table 4 reports simulation results for DGP3, which features groups with weak and strong first-stage effects that are not well-separated. For DGP2 and DGP3, we redefine the 2SLS-INF estimator as the infeasible estimator that chooses δ to minimize the theoretical MSE of the split-sample select-and-interact estimator stated in Thereom 2 using oracle information of ρ_g. The results in Tables 3 and 4 display similar patterns as those in Table 2. Both 2SLS-ADPT and 2SLS-SSL, designed to learn first-stage identification structure, out-perform the 2SLS-P, 2SLS-INT and 2SLS-SSINT estimators. Among its competitors, the proposed 2SLS-ADPT consistently performs the best or equally well.

Table 2: Rejection Proportion and MSE Performance for DGP 1

		2SLS-P	2SLS-INT	2SLS-SSINT	2SLS-INF	2SLS-ADPT	LIML-INT	2SLS-SSL	JIVE
$p_s = 0.05$, $\rho_{u,v} = 0.25$ and normal errors									
G = 40	$N\times$MSE	436.882	21.215	22.669	20.390	20.390	20.837	21.579	21.210
	$N\times$MAE	1923.779	443.467	423.261	430.740	430.740	432.387	426.859	435.797
	Rej. Prop.	0.038	0.066	0.046	0.050	0.050	0.058	0.050	0.040
G = 100	$N\times$MSE	458.475	23.663	20.486	19.609	19.599	19.740	19.808	19.807
	$N\times$MAE	3318.415	761.088	646.376	627.586	627.586	593.404	625.040	619.632
	Rej. Prop.	0.058	0.082	0.038	0.048	0.048	0.052	0.048	0.030
G = 200	$N\times$MSE	387.056	32.099	23.843	21.686	21.683	22.272	22.479	22.351
	$N\times$MAE	4341.239	1267.731	1043.499	1021.058	1018.517	1001.641	1007.512	1005.655
	Rej. Prop.	0.038	0.150	0.058	0.056	0.056	0.068	0.064	0.018
$p_s = 0.25$, $\rho_{u,v} = 0.25$ and normal errors									
G = 40	$N\times$MSE	16.240	4.221	4.334	4.215	4.215	4.223	4.253	4.238
	$N\times$MAE	385.748	206.675	202.563	198.870	198.870	203.090	196.234	200.877
	Rej. Prop.	0.054	0.042	0.042	0.042	0.042	0.048	0.042	0.048
G = 100	$N\times$MSE	17.488	4.594	4.390	4.399	4.399	4.365	4.374	4.359
	$N\times$MAE	661.972	311.558	306.842	320.676	320.676	307.353	322.333	307.540
	Rej. Prop.	0.062	0.076	0.066	0.072	0.072	0.078	0.068	0.050
G = 200	$N\times$MSE	15.208	4.578	4.156	4.051	4.050	4.046	4.093	4.058
	$N\times$MAE	875.906	469.002	427.221	421.565	421.565	418.238	425.306	410.839
	Rej. Prop.	0.042	0.068	0.056	0.050	0.050	0.058	0.052	0.034
$p_s = 0.05$, $\rho_{u,v} = 0.25$ and χ^2_3 errors									
G = 40	$N\times$MSE	6113.771	135.939	199.797	129.725	136.969	157.074	153.161	171.311
	$N\times$MAE	5278.398	1145.955	1321.611	1048.631	1052.631	1127.648	1156.241	1186.001
	Rej. Prop.	0.028	0.108	0.042	0.054	0.056	0.082	0.058	0.030
G = 100	$N\times$MSE	2791.348	189.829	191.282	124.294	142.841	146.779	143.370	155.182
	$N\times$MAE	7135.788	2340.853	1928.876	1541.093	1632.401	1796.948	1658.365	1833.990
	Rej. Prop.	0.026	0.174	0.054	0.058	0.064	0.074	0.054	0.014
G = 200	$N\times$MSE	2663.635	314.390	183.905	121.197	132.861	148.093	137.191	151.503
	$N\times$MAE	10743.992	4569.327	2901.436	2378.734	2545.821	2497.460	2494.414	2533.932
	Rej. Prop.	0.040	0.302	0.052	0.040	0.038	0.068	0.046	0.012
$p_s = 0.25$, $\rho_{u,v} = 0.25$ and χ^2_3 errors									
G = 40	$N\times$MSE	106.285	25.804	26.797	25.590	25.747	25.776	26.262	26.081
	$N\times$MAE	1072.737	492.951	532.019	496.556	501.732	486.464	532.610	504.988
	Rej. Prop.	0.050	0.068	0.042	0.050	0.054	0.060	0.054	0.036
G = 100	$N\times$MSE	94.806	30.146	29.370	26.310	27.498	26.725	27.206	27.058
	$N\times$MAE	1439.007	799.533	803.419	813.880	790.668	792.318	776.718	780.011
	Rej. Prop.	0.040	0.080	0.058	0.058	0.060	0.064	0.060	0.028
G = 200	$N\times$MSE	98.786	34.913	27.915	25.958	26.671	26.495	26.435	26.654
	$N\times$MAE	2165.718	1327.275	1159.424	1110.788	1140.030	1123.004	1140.352	1127.452
	Rej. Prop.	0.048	0.112	0.048	0.046	0.048	0.046	0.046	0.022

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.

Table 3: Rejection Proportion and MSE Performance for DGP 2

		2SLS-P	2SLS-INT	2SLS-SSINT	2SLS-INF	2SLS-ADPT	LIML-INT	2SLS-SSL	JIVE
$p_s = p_w = 0.025$, $\rho_{u,v} = 0.25$ and normal errors									
G = 100	$N\times$MSE	1435.660	43.079	47.703	40.468	41.944	42.666	44.084	43.596
	$N\times$MAE	3239.714	621.016	661.890	610.296	623.647	603.854	615.128	616.694
	Rej. Prop.	0.026	0.076	0.058	0.048	0.048	0.064	0.056	0.036
G = 100	$N\times$MSE	1905.386	64.325	54.582	46.870	48.225	49.529	50.044	50.688
	$N\times$MAE	6199.708	1318.320	1069.955	1036.453	1017.512	1016.949	1041.127	1022.186
	Rej. Prop.	0.038	0.118	0.050	0.046	0.042	0.070	0.056	0.020
G = 200	$N\times$MSE	1120.562	75.823	47.525	40.191	40.989	42.814	42.847	42.988
	$N\times$MAE	7228.982	2043.003	1493.240	1459.994	1445.584	1477.341	1478.735	1462.110
	Rej. Prop.	0.032	0.202	0.064	0.054	0.050	0.066	0.060	0.012
$p_s = p_w = 0.125$, $\rho_{u,v} = 0.25$ and normal errors									
G = 40	$N\times$MSE	45.407	8.481	8.708	8.414	8.546	8.391	8.575	8.413
	$N\times$MAE	647.919	265.315	276.036	260.195	272.046	261.139	270.705	260.026
	Rej. Prop.	0.054	0.076	0.062	0.054	0.056	0.050	0.058	0.046
G = 100	$N\times$MSE	51.337	9.325	8.816	8.750	9.021	8.658	8.706	8.635
	$N\times$MAE	1145.194	464.062	431.866	431.265	426.228	428.505	429.941	431.041
	Rej. Prop.	0.062	0.062	0.064	0.058	0.066	0.070	0.060	0.046
G = 200	$N\times$MSE	42.263	9.673	8.297	7.873	8.114	7.938	8.003	7.989
	$N\times$MAE	1456.95	673.573	604.213	582.558	565.383	570.043	583.826	581.422
	Rej. Prop.	0.040	0.080	0.050	0.054	0.052	0.046	0.050	0.034
$p_s = p_w = 0.025$, $\rho_{u,v} = 0.25$ and χ^2_3 errors									
G = 40	$N\times$MSE	720188.116	260.451	564.523	246.761	259.075	356.072	342.620	413.830
	$N\times$MAE	8722.373	1709.158	2091.322	1523.892	1644.423	1703.432	1842.463	1712.082
	Rej. Prop.	0.014	0.142	0.026	0.032	0.032	0.082	0.026	0.028
G = 100	$N\times$MSE	98979469.114	543.929	729.808	293.429	369.645	429.374	399.998	471.112
	$N\times$MAE	14080.389	4194.958	3740.958	2470.492	2619.759	2883.843	2789.741	2884.065
	Rej. Prop.	0.016	0.296	0.054	0.050	0.048	0.100	0.042	0.036
G = 200	$N\times$MSE	5137.888	586.768	356.761	187.578	220.009	253.785	229.412	262.706
	$N\times$MAE	14205.318	6525.808	3756.477	3003.463	3115.503	3338.126	3132.682	3373.438
	Rej. Prop.	0.038	0.420	0.058	0.040	0.036	0.090	0.048	0.012
$p_s = p_w = 0.125$, $\rho_{u,v} = 0.25$ and χ^2_3 errors									
G = 40	$N\times$MSE	303.819	51.212	59.449	52.709	54.365	52.746	55.503	54.548
	$N\times$ASE	1761.756	716.053	754.384	710.452	697.723	731.689	699.147	732.123
	Rej. Prop.	0.048	0.080	0.042	0.048	0.054	0.062	0.058	0.036
G = 100	$N\times$MSE	279.308	68.797	67.898	57.511	60.779	58.803	60.541	60.508
	$N\times$ASE	2458.124	1280.568	1188.038	1107.445	1163.068	1127.071	1093.128	1127.870
	Rej. Prop.	0.040	0.116	0.068	0.064	0.078	0.078	0.072	0.030
G = 200	$N\times$MSE	275.248	84.233	59.538	51.447	54.105	53.604	53.655	54.091
	$N\times$ASE	3568.601	2127.199	1583.559	1398.683	1518.890	1544.208	1528.103	1547.351
	Rej. Prop.	0.048	0.178	0.060	0.050	0.056	0.062	0.058	0.006

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.

Table 4: Rejection Proportion and MSE Performance for DGP 3

		2SLS-P	2SLS-INT	2SLS-SSINT	2SLS-INF	2SLS-ADPT	LIML-INT	2SLS-SSL	JIVE
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	34.822	35.869	35.534
	$N\times$MAE	1992.620	581.345	552.848	528.262	556.787	530.137	562.163	547.722
	Rej. Prop.	0.040	0.074	0.064	0.046	0.052	0.062	0.058	0.042
G = 100	$N\times$MSE	390.835	25.589	21.891	20.821	21.321	21.037	20.999	21.114
	$N\times$MAE	3071.019	786.978	686.456	648.956	668.154	632.800	671.309	652.069
	Rej. Prop.	0.058	0.084	0.046	0.054	0.054	0.052	0.052	0.032
G = 200	$N\times$MSE	316.284	33.967	24.779	22.619	23.071	23.142	23.374	23.229
	$N\times$MAE	3923.494	1328.891	1061.998	1039.967	1033.464	1061.145	1041.426	1079.376
	Rej. Prop.	0.038	0.146	0.068	0.052	0.056	0.060	0.048	0.028
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	270.101	290.787	311.778
	$N\times$MAE	5459.984	1440.192	1897.443	1441.298	1512.635	1517.961	1627.616	1531.046
	Rej. Prop.	0.026	0.134	0.028	0.042	0.050	0.086	0.040	0.034
G = 100	$N\times$MSE	2318.217	201.848	204.249	132.450	158.688	156.942	158.259	166.689
	$N\times$MAE	6620.035	2313.303	1980.954	1658.628	1879.294	1876.630	1761.201	1877.781
	Rej. Prop.	0.026	0.186	0.052	0.048	0.056	0.084	0.054	0.010
G = 200	$N\times$MSE	2150.827	327.106	190.661	125.153	138.088	153.325	143.897	156.877
	$N\times$MAE	9760.407	4617.979	2970.616	2480.836	2663.850	2604.448	2496.421	2702.159
	Rej. Prop.	0.042	0.308	0.054	0.030	0.030	0.080	0.050	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.

5. Empirical Examples

5.1. Return to Compulsory Schooling

The return to schooling literature studies how an extra year of schooling affects individual outcomes, such as earnings and health outcomes, later in life. 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 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 with indicators for geographic regions and demographic groups. The dataset include 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

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

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_{st} is the group indicator which varies with birth cohort t and birth state s. The exogenous regressor X_{ist} includes survey year, birth state, census division by gender and race, and census division by birth year fixed effects, and a fourth-order polynomial in age when applicable (as we explain 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.

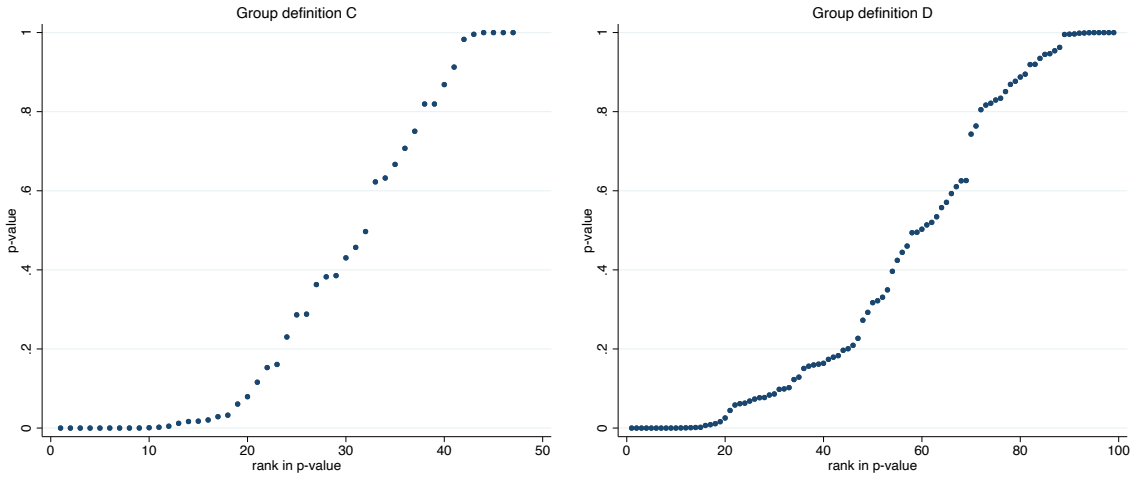
Besides allowing for first-stage heterogeneity, our simultaneous equation model is the same as the one used in Table 1 of Stephens and Yang (2014), except that Stephens and Yang (2014) uses three indicators (corresponding to being required to attend seven, eight, and nine or more years of schooling) constructed from CL_{st} as instruments while we use CL_{st} directly. We adopt this specification because our formal results consider only the scalar IV case. In addition, our specification includes census division by year-of-birth fixed effects, while Stephens and Yang (2014) include census region by year-of-birth fixed effects. 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 the heterogeneity in the enforcement of compulsory schooling laws and in school quality across the U.S. historically. Gender and race are used because the literature has found that males are most responsive to changes in the minimum school leaving age and often selects them for subsample analysis. Besides males, Oreopoulos (2006) also selects non-White males for subsample analysis, while Stephens and Yang (2014) selects White males. Survey years are used because of concerns in survey accuracy in earlier years. When survey year is used to define groups, the fourth-order polynomial in age is omitted from X_{ist} as age is then perfectly collinear with birth year fixed effects in groupwise regressions.

Figure 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 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.

Figure 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.

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, as the demographic control. 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 results are not reported in panels A1-B1 and A2-B2 because both *Stata* and *R* fail to compute LIML with these model specifications due to multicollinearity across census division by birth year fixed effects, census year indicators, and a fourth-order polynomial in age. LIML 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.

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 simulations. 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 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.

Table 5: Return to Compulsory Schooling: Estimation Results

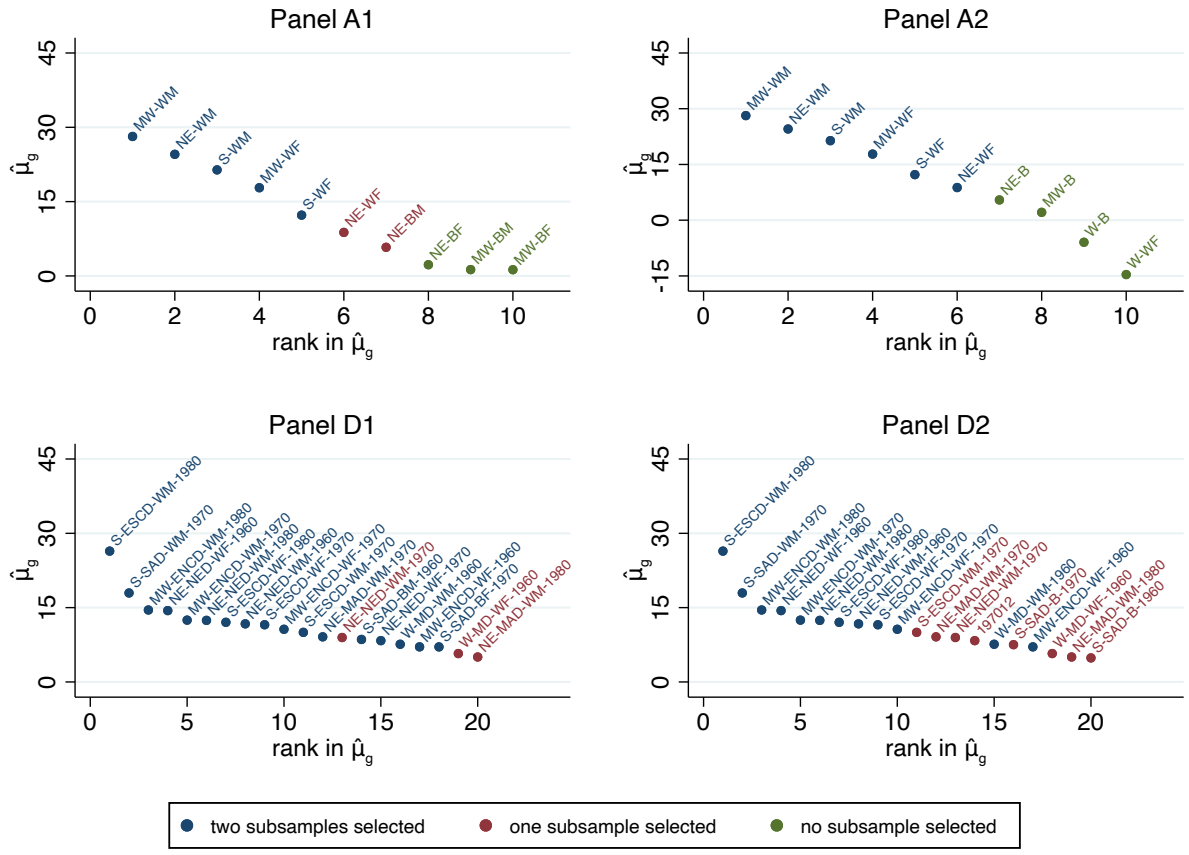
Full-sample				Select-and-interact			
OLS	2SLS-P	2SLS-INT	LIML-INT	2SLS-SSL	2SLS-ADPT		
					(κ^*)	($2\kappa^*$)	($\kappa^*/2$)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panels A1-D1: four gender and race categories							
Panel A1: groups defined by census region, gender, and race							
0.070***	-0.254	0.066***	-	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 uses four demographic groups: White males, White females, non-White males, and non-White females. Panels A1-D1 uses 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.

The graphs in Figure 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 correspond to Panel A1 and A2 in Table 5, which are the coarsest definition of groups we've considered. The bottom two graphs correspond to Panel D1 and D2, which are the finest definition of groups we've considered. Each dot in the graphs represents a group.

The selection of two, one, or none of the subsamples by our adaptive 2SLS procedure is color-coded. 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 and groups in West divisions do not seem to contribute much to identification.

Figure 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.

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, thereafter) and coal supply. This strategy follows the earlier work by Black, Daniel, and Sanders (2002), who utilizes coal shocks to study the impact of local economic conditions on participation in programs of disability payments, and Acemoglu, Finkelstein, and Notowidigdo (2013) who utilizes oil shocks to study the effect of local income on health spending. Recently, Charles, Li, and Stephens (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 first-stage 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 to be Kentucky, Ohio, Pennsylvania, and West Virginia, following Black, Daniel, and Sanders (2002), and defines oil states to be 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, Li, and Stephens (2018) adds Louisiana to the list of oil states. Acemoglu, Finkelstein, and Notowidigdo (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. Let subscript c denote county, s denote state, and t denote year (when an election takes

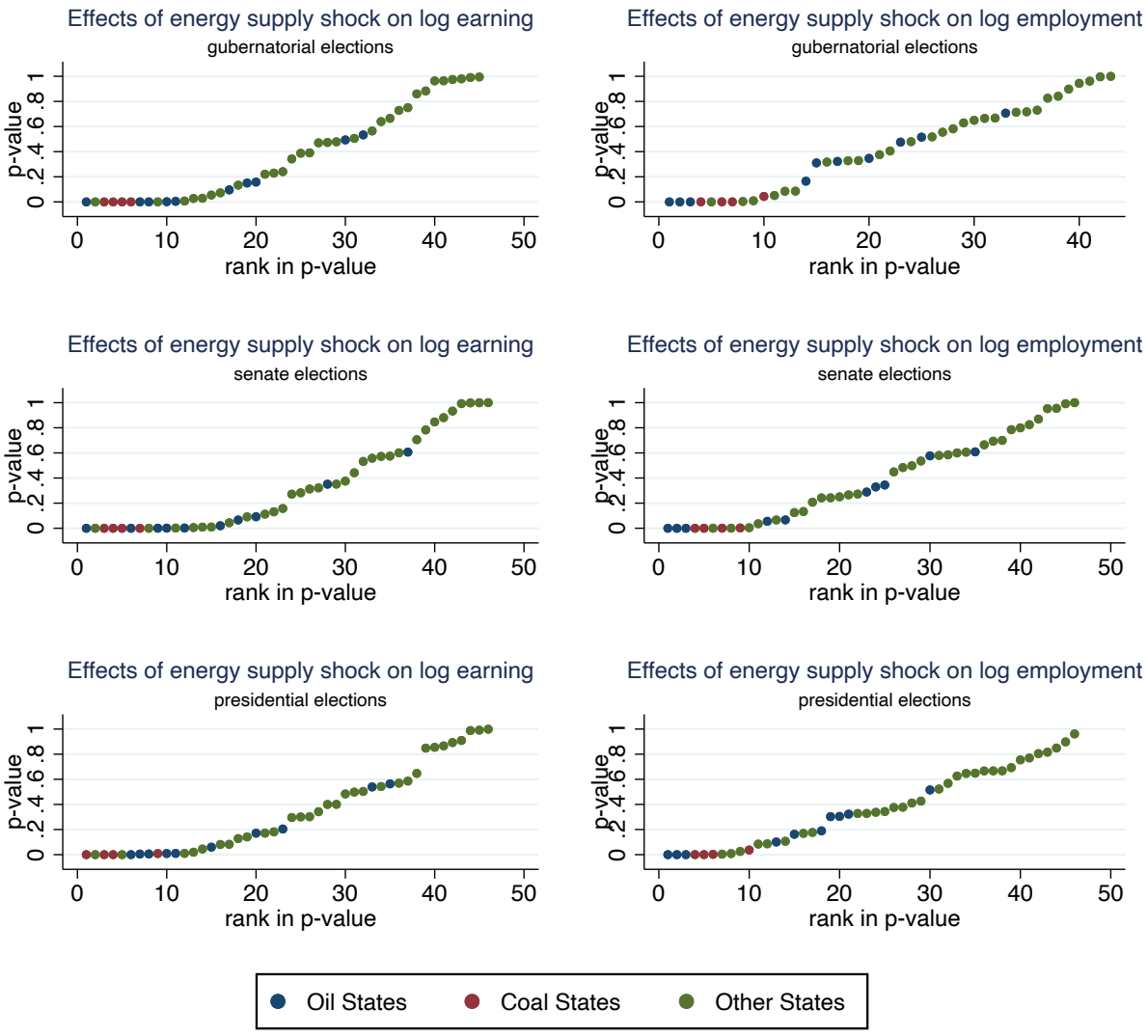
place). We consider the model,

$$\begin{aligned} \Delta Vote_{cst} &= \beta \Delta Economy_{cst} + X_{cst}\theta_s + u_{cst} \\ \Delta Economy_{cst} &= \sum_{s=1}^{S} \rho_s \Delta EnergySupply_t \times EnergyShare_{cs} + X_{cst}\gamma_s + v_{cst}, \end{aligned} \tag{7}$$

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 documented in 1967 CBP, and X_{cst} is the list of exogenous regressors in Charles and Stephens (2013), which includes county-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. Indeed, the use of the 1967 employment measure instead of the 1974 measure to construct the instruments of the model in equation (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 employment share in the mining industry. As we show below, β becomes significant when it is estimated with our adaptive procedure.

Figure 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.

Graphs in Figure 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.

Furthermore, 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, Figure 4 provides ample motivation to apply our proposed methodology of selecting strong first-stage signals with the target of minimizing the asymptotic MSE.

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.

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 (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. This finding is qualitatively the same as the results reported in Charles and Stephens (2013), who find that higher local

wages and employment lower turnout in elections for Governor and Senator but have no effect on presidential turnout.

All estimators reported in columns (3)-(8) are first-order equivalent under the assumptions discussed 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 result 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 a interacted 2SLS preserves first-order efficiency but may still have substantial higher-order asymptotic bias. Following earlier work of 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. 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 lasso of Belloni, Chen, Chernozhukov, and Hansen (2012) applied to the case when the first stage regressors are interactions between an instrument 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 lasso estimator, because both methods consistently select all groups with strong instruments in the first-stage. When weak instrument proportion does not go to zero, both estimators are consistent and asymptotic normal. To the best of our knowledge, there is no higher order analysis of the split-sample lasso estimator. In simulations, we find that our proposed adaptive estimator behaves comparably and in some DGPs better than the split-sample lasso selection estimator.

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 allows us to replicate the main results of Charles and Stephens (2013) using an alternative sample with a weaker instrument, but with higher plausibility of the exclusion restriction.

Appendix A: Proofs of Auxiliary Lemmas

In the appendix, we replace $c_{g,\alpha_{FS}}$ by c_g and α_{FS} by α for convenience in notations. We will use $\lesssim$ to denote that an inequality holds up to a universal constant for all groups. For instance, a random element $A_g \lesssim n_g$ means that there exists a universal constant $\mathcal{C} < \infty$ such that $A_g/n_g \leq \mathcal{C}$ for large enough n_g, for all $g = 1, ..., G$. Also, let $\tilde{X}_g = [Z_g \ X_g]$ be an $n_g \times (d+1)$ matrix. For all $g = 1, ..., G$, let $H_{g,1} = \sqrt{W_g' M_{\tilde{X}_g} W_g / n_g}$ and $H_{g,2} = \sqrt{Z_g' Z_g / n_g} = \sqrt{\tilde{Z}_g' M_{X_g} \tilde{Z}_g / n_g}$.

Lemma A1. *Suppose we use Bonforroni-type correction to simultaneously test $H_0 : \rho_g = 0$ vs $H_a : \rho_g > 0$ for all $g = 1, 2, \ldots, G$. The implied thresholding value δ falls inside the δ range specified in Assumption 2 as long as $G \log G/N \to 0$ as $G, N \to \infty$.*

Proof. The threshold δ^* corresponding to the Bonferroni-type multiple testing controlling family-wise error rate α test would satisfy that $1 - \Phi(\delta^*/\sigma_v) = \frac{\alpha}{G}$. Without loss of generality, set $\sigma_v = 1$. Since $1 - \Phi(x) \leq \phi(x)/x$ for all $x > 0$, we know that as long as $1 - \frac{\alpha}{G} > 0.5$, $\frac{\alpha}{G} \leq \phi(\delta^*)/\delta^*$. For any $\delta^* > 1$, we further have that $\frac{\alpha}{G} \leq \phi(\delta^*) \leq \exp(-(\delta^*)^2/2)$. Therefore, $\delta^* = O(\sqrt{\log G})$, which implies that $\delta^* = o((\frac{N}{G})^{1/2})$ as long as $G \log G/N \to 0$ as $G, N \to \infty$. Other multiple testing procedures are less stringent than the Bonferroni correction (see a comparison in Genovese and Wasserman, 2002), hence the associated threshold δ corresponding to those procedures will not be larger than the Bonferroni method. □

Lemma A2. *Under Assumption 1,*

$$P\left(|H_{g,1}^2 - \sigma_v^2| > \frac{\sigma_v^2}{2}\right) \lesssim \frac{1}{n_g^2}, \quad P\left(|H_{g,2}^2 - k_g| > \frac{k_g}{2}\right) \lesssim \frac{1}{n_g^2}.$$

Proof. Notice that $H_{g,1}^2 = W_g' M_{\tilde{X}_g} W_g / n_g = v_g' M_{\tilde{X}_g} v_g / n_g = v_g' v_g / n_g - v_g' P_{\tilde{X}_g} v_g / n_g$, where $E[v_g' P_{\tilde{X}_g} v_g | \tilde{X}_g] = tr(P_{\tilde{X}_g}) E[v_g v_g' | \tilde{X}_g] = \sigma_v^2 (d+1)$. We know that for large enough n

$$\begin{aligned}
P(|H_{g,1}^2 - \sigma_v^2| > \sigma_v^2/2) \leq & P(|v_g' v_g/n_g - \sigma_v^2| + v_g' P_{\tilde{X}_g} v_g/n_g > \sigma_v^2/2) \\
\leq & P(|v_g' v_g/n_g - \sigma_v^2| > \sigma_v^2/4) + P(v_g' P_{\tilde{X}_g} v_g/n_g > \sigma_v^2/4) \\
\leq & P(|v_g' v_g/n_g - \sigma_v^2| > \sigma_v^2/4) + P(v_g' P_{\tilde{X}_g} v_g/n_g - \sigma_v^2(d+1)/n_g > \sigma_v^2/8) \\
\leq & E[(v_g' v_g/n_g - \sigma_v^2)^4]/(\sigma_v^2/4)^4 + E[(v_g' P_{\tilde{X}_g} v_g)^2]/n_g^2/(\sigma_v^2/8))^2,
\end{aligned}$$

where

$$E\left[\left(v_g'v_g/n_g-\sigma_v^2\right)^4\right]=\left[V\left(v_g'v_g/n_g\right)\right]^2+V\left[\left(v_g'v_g/n_g-\sigma_v^2\right)^2\right]\lesssim 1/n_g^2,$$

uniformly over all g due to the moment condition of the error term v_{ig} in Assumption 1. To bound $E[(v_g'P_{\tilde{X}_g}v_g)^2]$ uniformly over all g, let P_{ij} denote the (i,j)-th element of $P_{\tilde{X}_g}$ and notice that $\sum_{i=1}^{n_g}P_{ii}=tr(P_{\tilde{X}_g})=d+1$, $0\leq P_{ii}\leq 1$. It follows that $0\leq\sum_{i\neq j}P_{ii}P_{jj},\sum_i(P_{ii})^2\leq(\sum_i P_{ii})^2=(d+1)^2$, and $0\leq\sum_{i\neq j}P_{ij}P_{ij}\leq\sum_{i,j}P_{ij}^2=tr(P_{\tilde{X}_g}'P_{\tilde{X}_g})=d+1$. Then by the moment condition of the error term v_{ig} in Assumption 1, we know

$$\begin{aligned}E[(v_g'P_{\tilde{X}_g}v_g)^2|X_g]&=\sum_{i,j,k,l}E[v_{ig}P_{ij}v_{jg}v_{kg}P_{kl}v_{lg}|X_g]\\&=\sum_i P_{ii}^2E(v_{ig}^4|X_g)+\sum_{i\neq j}P_{ii}P_{jj}E(v_{ig}^2v_{jg}^2|X_g)+\sum_{i\neq j}(P_{ij}^2+P_{ij}P_{ji})E(v_{ig}^2v_{jg}^2|X_g)\end{aligned}$$

is bounded by a universal constant across all groups.

The second inequality for $H_{g,2}^2=\tilde{Z}_g'M_{X_g}\tilde{Z}_g/n_g=\eta_g'M_{X_g}\eta_g/n_g$ could be proven with the same arguments as above. We omit the details. □

Lemma A3. *Under Assumption 1 and provided that $\delta\leq C_\delta(\frac{N}{G})^{1/2}$ with $C_\delta<\underline{\rho}\sqrt{\underline{kc}/2}$,*

$$\sup_{\delta\leq C_\delta(\frac{N}{G})^{1/2}}P(\hat{\mu}_g\leq\delta)\lesssim\frac{1}{n_g^2}.$$

Proof. We only need to show that the above rate condition holds for all $g\in G_{+,s}$ as the results for other groups follow from that of the strong group. Use $\sup_\delta$ in short for $\sup_{\delta\leq C_\delta(\frac{N}{G})^{1/2}}$. Since $C_\delta<\underline{\rho}\sqrt{\underline{kc}/2}$, then there exists a small positive constant $\eta\in(0,1)$ such that $C_\delta\leq$

$\underline{\rho}\sqrt{\underline{kc}/2}(1-\eta)$. Given that η, we know

$$\begin{aligned}
&\sup_{\delta} P\left(\hat{\mu}_g \leq \delta\right) \leq \sup_{\delta} P\left(\delta > \sqrt{1-\eta}\cdot\mu_g\right) + \sup_{\delta} P\left(\hat{\mu}_g \leq \sqrt{1-\eta}\cdot\mu_g\right) \\
\leq &\sup_{\delta} P\left(\sqrt{Z_g'Z_g} < \delta/\rho_g/\sqrt{1-\eta}\right) + P\left(Z_g'v_g/\sqrt{Z_g'Z_g} \leq (\sqrt{1-\eta}-1)\cdot\mu_g\right) \\
\leq &\sup_{\delta} P\left(\sqrt{Z_g'Z_g} < \sqrt{(1-\eta)\underline{kc}/2}\cdot\sqrt{N/G}\right) \\
&+ P\left(Z_g'v_g/\sqrt{Z_g'Z_g} \leq (\sqrt{1-\eta}-1)\rho_g\sqrt{\frac{3}{2}k_g n_g}\right) + P(Z_g'Z_g/n_g > \frac{3}{2}k_g) \\
\leq &P\left(H_{g,2}^2 < (1-\eta)k_g\right) + P\left(v_g'P_{Z_g}v_g/n_g \geq \frac{3}{2}(\sqrt{1-\eta}-1)^2\underline{\rho}^2\underline{k}\right) + P\left(H_{g,2}^2 > \frac{3}{2}k_g\right) \\
\lesssim &\frac{1}{n_g^2}.
\end{aligned}$$

The last line follows since, with similar arguments as those in Lemma A2, one can show that $P\left(|H_{g,2}^2 - k_g| > \eta\cdot k_g\right) \lesssim \frac{1}{n_g^2}$ and $P\left(v_g'P_{Z_g}v_g/n_g \geq C\right) \lesssim \frac{1}{n_g^2}$ for any positive η and C, respectively. The lemma is then proven. □

Lemma A4. *Under Assumptions 1 and 2, for any non-negative integer k, as $G, N \to \infty$,*

1. $\sup_{\delta \leq C_\delta(\frac{N}{G})^{1/2}} |(\delta-\mu_g)/\sigma_v|^k \phi((\delta-\mu_g)/\sigma_v) < \infty$ *for all g;*

2. $E\Big[|(\delta-\mu_g)/\sigma_v|^k\phi((\delta-\mu_g)/\sigma_v)\Big] \lesssim \frac{1}{n_g^2}$ *and* $E\Big[\Phi((\delta-\mu_g)/\sigma_v)\Big] \lesssim \frac{1}{n_g^2}$ *for all $g \in \mathcal{G}_{+,s}$;*

3. $\sup_{\delta \leq C_\delta(\frac{N}{G})^{1/2}} \sum_{g\in\mathcal{G}_{+,w}} \mu_g^h |(\delta-\mu_g)/\sigma_v|^k \phi((\delta-\mu_g)/\sigma_v)/N = O_p(G/N)$,
   $\sup_{\delta \leq C_\delta(\frac{N}{G})^{1/2}} \sum_{g\in\mathcal{G}_{+,w}} \mu_g^h \Phi((\delta-\mu_g)/\sigma_v)/N = O_p(G/N)$ *for $h = 1, 2$;*

4. $\sup_{\delta \leq C_\delta(\frac{N}{G})^{1/2}} \sum_{g\in\mathcal{G}_{+,s}} \mu_g |(\delta-\mu_g)/\sigma_v|^k \phi((\delta-\mu_g)/\sigma_v)/N = O_p(G^2/N^2) = o_p(G/N)$,
   $\sup_{\delta \leq C_\delta(\frac{N}{G})^{1/2}} \sum_{g\in\mathcal{G}_{+,s}} \mu_g \Phi((\delta-\mu_g)/\sigma_v)/N = O_p(G^2/N^2) = o_p(G/N)$,
   $\sup_{\delta \leq C_\delta(\frac{N}{G})^{1/2}} \sum_{g\in\mathcal{G}_{+,s}} \mu_g^2 |(\delta-\mu_g)/\sigma_v|^k \phi((\delta-\mu_g)/\sigma_v)/N = O_p((G/N)^{3/2}) = o_p(G/N)$ *and*
   $\sup_{\delta \leq C_\delta(\frac{N}{G})^{1/2}} \sum_{g\in\mathcal{G}_{+,s}} \mu_g^2 \Phi((\delta-\mu_g)/\sigma_v)/N = O_p((G/N)^{3/2}) = o_p(G/N)$.

Proof. In this proof, $\sup_\delta$ is used in short for $\sup_{\delta \leq C_\delta(\frac{N}{G})^{1/2}}$. For the first statement, notice that the exponential function has the property that for any $x, l > 0$, $e^x \geq 1 + x^l/l!$. That is,

$e^{-x^2} \leq l!/x^{2l}$. Therefore,

$$|(\delta-\mu_g)/\sigma_v|^k \phi((\delta-\mu_g)/\sigma_v) \leq 1/\sqrt{2\pi} \cdot |(\delta-\mu_g)/\sigma_v|^k \cdot (k/2)!/(|(\delta-\mu_g)/\sigma_v|)^k = (k/2)!/\sqrt{2\pi},$$

and hence the first statement holds.

For the second statement, notice that for any non-negative integer k, we have

$$\begin{aligned}
&\sup_\delta E\Big[|(\delta-\mu_g)/\sigma_v|^k e^{-((\delta-\mu_g)/\sigma_v)^2}\Big] \leq \sup_\delta E\Big[e^{k\cdot(|\delta-\mu_g|/\sigma_v-1)-((\delta-\mu_g)/\sigma_v)^2}\Big] \\
&= e^{k^2/4-k} \sup_\delta E\Big[e^{-(|\delta-\mu_g|/\sigma_v-k/2)^2}\Big] \leq e^{k^2/4-k} \sup_\delta \zeta_g + e^{k^2/4-k} \sup_\delta P(e^{-(|\delta-\mu_g|/\sigma_v-k/2)^2} > \zeta_g) \\
\leq& e^{k^2/4-k}\zeta_g + e^{k^2/4-k} \sup_\delta P(e^{-(1-\sqrt{1-\eta})^2\mu_g^2/\sigma_v^2/2} > \zeta_g) + e^{k^2/4-k} \sup_\delta P\Big(\delta \geq \sqrt{1-\eta}\cdot\mu_g\Big),
\end{aligned}$$

where the last inequality holds for large enough n_g as both k and η are fixed. Set $\zeta_g = e^{-n_g k_g (1-\sqrt{1-\eta})^2\rho_g^2/\sigma_v^2/8}$, we know that $\zeta_g \lesssim \frac{1}{n_g^2}$, and $P(e^{-(1-\sqrt{1-\eta})^2\mu_g^2/\sigma_v^2/2} > \zeta_g) = P(Z_g'Z_g < k_g n_g/4) \lesssim \frac{1}{n_g^2}$. The first inequality in the second statement is then proven as $\sup_\delta P\Big(\delta \geq \sqrt{1-\eta}\cdot\mu_g\Big) \lesssim \frac{1}{n_g^2}$ is already shown at the end of the proof for Lemma A3.

Denote $\Phi((\delta-\mu_g)/\sigma_v)$ by $\Phi_{g,\delta}$ for simplicity. For the second inequality in the second statement of the lemma, notice that

$$\begin{aligned}
&\sup_\delta E\Big[\Phi_{g,\delta}\Big] = \sup_\delta \Big\{E\Big[\Phi_{g,\delta}1\Big(\delta < \sqrt{1-\eta}\mu_g\Big)\Big] + E\Big[\Phi_{g,\delta}1\Big(\delta \geq \sqrt{1-\eta}\mu_g\Big)\Big]\Big\} \\
\leq& E\Big[\phi\Big(\frac{(1-\sqrt{1-\eta})\mu_g}{\sigma_v}\Big)/\frac{(1-\sqrt{1-\eta})\mu_g}{\sigma_v}\Big] + \sup_\delta P(\delta \geq \sqrt{1-\eta}\cdot\mu_g),
\end{aligned}$$

where the inequality holds as $\Phi(-d) \leq \phi(d)/d$ for all $d > 0$. Following the same arguments as in proving the first inequality of the second statement, $E\Big[\phi\Big(\frac{(1-\sqrt{1-\eta})\mu_g}{\sigma_v}\Big)/\frac{(1-\sqrt{1-\eta})\mu_g}{\sigma_v}\Big] \lesssim \frac{1}{n_g^2}$. Following the proof of Lemma A3, $\sup_\delta P\Big(\delta \geq \sqrt{1-\eta}\cdot\mu_g\Big) \lesssim \frac{1}{n_g^2}$.

The third statement of the lemma is apparent given the first statement and the fact that $\sum_{g\in\mathcal{G}_{+,w}} \mu_g^2/N \leq \bar{\rho}^2 \cdot \frac{1}{G_{+,w}} \sum_{g\in\mathcal{G}_{+,w}} Z_g'Z_g/n_g \cdot \frac{G_{+,w}}{N} = O_p(G/N)$, and $\sum_{g\in\mathcal{G}_{+,w}} \mu_g/N \leq \bar{\rho} \cdot \frac{1}{G_{+,w}} \sum_{g\in\mathcal{G}_{+,w}} \sqrt{Z_g'Z_g/n_g} \cdot \frac{G_{+,w}}{N} = O_p(G/N)$, where all the convergence results hold by LLN for inid sequences.

Denote $\phi((\delta-\mu_g)/\sigma_v)$ by $\phi_{g,\delta}$ for simplicity. To show the first inequality of the last statement when $k = 0$, we notice that

$$\sup_\delta E[\mu_g\phi_{g,\delta}] \leq \sup_\delta \sqrt{E[\mu_g^2\phi_{g,\delta}]}\sqrt{E[\phi_{g,\delta}]} \leq \bar{\rho} E[(Z_g'Z_g)^2]^{1/4} \sup_\delta E[\phi_{g,\delta}^2]^{1/4} \sup_\delta E[\phi_{g,\delta}]^{1/2} \lesssim 1/n_g$$

where follows the same arguments as in proving the second statement of the lemma one can show that $\sup_\delta E[\phi_{g,\delta}^2] \lesssim 1/n_g^2$. Similarly, $\sup_\delta E[\mu_g^2\phi_{g,\delta}] \lesssim 1/\sqrt{n_g}$, $\sup_\delta E[\mu_g\Phi_{g,\delta}] \lesssim 1/n_g$, and $\sup_\delta E[\mu_g^2\Phi_{g,\delta}] \lesssim 1/\sqrt{n_g}$. The last statement is therefore proven. □

Lemma A5. *Under Assumption 1, we have that as $G, N \to \infty$, if the additional rate condition $G \log G/N \to 0$ holds, then (i) $\frac{N/G}{\min_{g\in\mathcal{G}_{+,s}} \mu_g^2} = O_p(1)$; (ii) if in addition $\tilde{Z}$ is assumed to be sub-exponential, we have $\max_{g\in\mathcal{G}_{+,w}} \mu_g^2 = O_p(1)$.*

Proof. To show (ii), we first note that

$$\max_{g\in\mathcal{G}_{+,w}} \mu_g^2 \leq \max_{g\in\mathcal{G}_{+,w}} a_g Z_g' Z_g/n_g \leq \bar{\rho} \max_{g\in\mathcal{G}_{+,w}} Z_g' Z_g/n_g.$$

Since $Z_g' Z_g \leq \tilde{Z}_g' \tilde{Z}_g$, then it suffices to prove that $\max_{g\in\mathcal{G}_{+,w}} \frac{1}{n_g} \sum_i \tilde{Z}_{ig}^2 = O_p(1)$ under the conditions stated. Under the assumption that $\tilde{Z}_{ig}$ is sub exponential, there exists some $\lambda > 0$ such that $\tilde{k}_g = \mathbb{E}[\exp(\lambda \tilde{Z}_{ig}^2)]$ exists for all $g \in \mathcal{G}_{+,w}$ and $\max_{g\in\mathcal{G}_{+,w}} \tilde{k}_g < \infty$. Note that if $\tilde{Z}_{ig}$ is a bounded random variable, which is commonly the case for an instrument (i.e. binary indicator or proportions), then the exponential moment existence condition is satisfied.

Under the stated condition, for any $\epsilon > 0$

$$\begin{aligned} P\left(\max_{g\in\mathcal{G}_{+,w}} \tilde{Z}_g' \tilde{Z}_g/n_g > \epsilon\right) &\leq \sum_{g\in\mathcal{G}_{+,w}} P\left(\sum_i \tilde{Z}_{ig}^2 > n_g\epsilon\right) = \sum_{g\in\mathcal{G}_{+,w}} P\left(\exp(\lambda \sum_i \tilde{Z}_{ig}^2) > \exp(\lambda n_g \epsilon)\right) \\ &\leq \sum_{g\in\mathcal{G}_{+,w}} \frac{E[\exp(\lambda \sum_i \tilde{Z}_{ig}^2)]}{\exp(\lambda n_g \epsilon)} = \sum_{g\in\mathcal{G}_{+,w}} \frac{\tilde{k}_g^{n_g}}{\exp(\lambda n_g \epsilon)} \\ &= \sum_{g\in\mathcal{G}_{+,w}} \exp\left(-n_g(\lambda\epsilon - \log \tilde{k}_g)\right). \end{aligned}$$

Now pick $\lambda\epsilon > \max_{g\in\mathcal{G}_{+,w}} \log \tilde{k}_g$, which is finite, then we can find a constant $\tau > 0$ such that

$$\begin{aligned} P\left(\max_{g\in\mathcal{G}_{+,w}} \tilde{Z}_g' \tilde{Z}_g/n_g > \epsilon\right) &\leq G_{+,w} \exp(-\frac{N}{G}\tau) \leq G \exp(-\frac{N}{G}\tau) \\ &= \exp\left(\frac{N}{G}\Big(\frac{G\log G}{N} - \tau\Big)\right) \to 0 \end{aligned}$$

provided $G \log G/N \to 0$.

To show (i), we note that we have $Z_{ig} \equiv \tilde{Z}_{ig} - X_{ig}\hat{\lambda}$ where $\hat{\lambda}$ is the OLS estimator for the coefficient λ in the linear model $\tilde{Z}_g = X_g\lambda + \eta_g$. For all $g \in \mathcal{G}_{+,s}$, we have that

$$\begin{aligned}\mu_g^2 &= \sum_{i=1}^{n_g} \rho_g^2 Z_{ig}^2 = \rho_g^2 \sum_{i=1}^{n_g} \Big(\tilde{Z}_{ig} - X_{ig}'\lambda + X_{ig}'\lambda - X_{ig}'\hat{\lambda}\Big)^2 \\ &\geq \rho_g^2 \sum_{i=1}^{n_g} \eta_{ig}^2 + 2\sum_{i=1}^{n_g} \eta_{ig} X_{ig}'(\lambda - \hat{\lambda}) \geq \bar{\rho}^2 \sum_{i=1}^{n_g} \eta_{ig}^2 + 2\sum_{i=1}^{n_g} \eta_{ig} X_{ig}'(\lambda - \hat{\lambda}).\end{aligned}$$

Then for any C,

$$P\left(\min_{g\in\mathcal{G}_{+,s}} \mu_g^2 \leq C\right) \leq P\left(\min_{g\in\mathcal{G}_{+,s}} \left(\bar{\rho}^2 \sum_{i=1}^{n_g} \eta_{ig}^2 + 2\sum_{i=1}^{n_g} \eta_{ig} X_{ig}'(\lambda - \hat{\lambda})\right) \leq C\right) \leq P\left(\min_{g\in\mathcal{G}_{+,s}} \bar{\rho}^2/2 \sum_{i=1}^{n_g} \eta_{ig}^2 \leq C\right).$$

The last inequality holds since $\sum_{i=1}^{n_g} \eta_{ig}^2 = O_p(n_g)$ and $\sum_{i=1}^{n_g} \eta_{ig} X_{ig}'(\lambda - \hat{\lambda}) = O_p(\sqrt{n_g} \cdot 1/\sqrt{n_g}) = O_p(1)$.

Lastly, using the one-sided Bernstein inequality positive random variables, we know that

$$P\Big(\sum_{i=1}^{n_g} \Big(\eta_{ig}^2 - k_g\Big) \leq -n_g k_g/2\Big) \leq \exp\Big(-\frac{n_g(k_g/2)^2}{2\mathbb{E}[\eta_{ig}^4]}\Big).$$

Under assumption 1, we know there exists some $\bar{\Delta}_Z$ such that $\mathbb{E}[\eta_{ig}^4] \leq \bar{\Delta}_Z$ for all $g = 1, 2, ..., G$. Then

$$P\Big(\sum_{i} \eta_{ig}^2 \leq \underline{k}/2 \cdot \underline{c}\frac{N}{G}\Big) \leq P\Big(\sum_{i=1}^{n_g} \Big(\eta_{ig}^2 - k_g\Big) \leq -n_g k_g/2\Big) \leq \exp\Big(-\frac{\underline{c}\underline{k}^2 \frac{N}{G}}{8\bar{\Delta}_z}\Big).$$

Therefore,

$$\begin{aligned}P\Big(\min_{g\in\mathcal{G}_{+,s}} \mu_g^2 \leq \frac{1}{4}\bar{\rho}^2 \underline{c}\underline{k}\frac{N}{G}\Big) &\leq \sum_{g\in\mathcal{G}_{+,s}} P\Big(\sum_{i} \eta_{ig}^2 \leq \underline{k}/2 \cdot \underline{c}\frac{N}{G}\Big) \leq G\exp\Big(-\frac{\underline{c}\underline{k}^2 \frac{N}{G}}{8\bar{\Delta}_z}\Big) \\ &= \exp\Big(\frac{N}{G}\Big(\frac{G\log G}{N} - \frac{\underline{c}\underline{k}^2}{8\bar{\Delta}_z}\Big)\Big) \to 0.\end{aligned}$$

The Lemma is proven.

□

Appendix B: Proof of Theorems and Corollaries

Proof of Some Limiting Results Stated in Section 2.2:

Asymptotic Results of $\hat{\beta}_{pool}$ and $\hat{\beta}_{int}$ in Section 2.2.1:

Proof. First, we note that the pooled estimator $\hat{\beta}_{pool} = (Z'W)^{-1}Z'Y$ defined in Section 2.2.1 using the groupwise transformed instrument Z is equivalent to a 2SLS estimator of β using the original excluded variable $\tilde{Z}$ as the instrument in a model that also controls for group-specific X in both stages. This is because

$$\begin{aligned}
&\left[\left((W\ D_X)'P_{(\tilde{Z}\ D_X)}(W\ D_X)\right)^{-1}(W\ D_X)'P_{(\tilde{Z}\ D_X)}Y\right]_1 \\
&= \left[\left((W\ D_X)'P_{(Z\ D_X)}(W\ D_X)\right)^{-1}(W\ D_X)'P_{(Z\ D_X)}Y\right]_1 \\
&= (Z'W)^{-1}Z'Y = \hat{\beta}_{pool},
\end{aligned}$$

where $[.]_1$ denotes the first element of a vector.

Similarly, the fully-interacted 2SLS estimator defined in Section 2.2.1 is equivalent to a 2SLS estimator of β in a model that uses group-specific $\tilde{Z}$ as the instrument and group-specific X as the exogenous control in both stages.

$$\begin{aligned}
\hat{\beta}_{int} &= \left[\left((W\ D_X)'P_{(\tilde{D}\ D_X)}(W\ D_\ell)\right)^{-1}(W\ D_\ell)'P_{(\tilde{D}\ D_X)}Y\right]_1 \\
&= \left[\left((W\ D_X)'P_{(D\ D_X)}(W\ D_X)\right)^{-1}(W\ D_X)'P_{(D\ D_X)}Y\right]_1 \\
&= (W'P_DW)^{-1}W'P_DY.
\end{aligned}$$

Next, we show the limiting results of the two estimators stated in Section 2.2.1.

The pooled estimator $\hat{\beta}_{pool} = (Z'W)^{-1}Z'Y = \beta + (Z'W)^{-1}Z'u$. The estimator is consistent because by the LLN for independent and not identically distributed sequences and the

rate condition $G/N \to 0$,

$$\begin{aligned} Z'W/N &= \sum_{g=1}^{G} Z_g'(Z_g\rho_g + X_g\omega_g + v_g)/N = \sum_{g=1}^{G} \rho_g Z_g' Z_g/n_g \cdot n_g/N + \sum_{g=1}^{G} Z_g' v_g/N \\ &= \sum_{g=1}^{G} \rho_g \tilde{Z}_g' M_{X_g} \tilde{Z}_g/n_g \cdot n_g/N + o_p(1) = \sum_{g=1}^{G} \rho_g \eta_g' \eta_g/n_g \cdot n_g/N - \sum_{g=1}^{G} \rho_g \eta_g' P_{X_g} \eta_g/N + o_p(1) \\ &= \sum_{g=1}^{G} \rho_g E[\eta_g'\eta_g/n_g] p_g + o_p(1) = \sum_{g=1}^{G} \rho_g k_g p_g + o_p(1); \\ Z'u/N &= \sum_{g=1}^{G} Z_g' u_g/N = o_p(1). \end{aligned}$$

In addition, since

$$\sum_{g=1}^{G} V[Z_g' u_g/\sqrt{n_g} \cdot \sqrt{n_g/N}] = \sum_{g=1}^{G} E[Z_g' u_g u_g' Z_g/n_g] p_g = \sigma_u^2 \sum_{g=1}^{G} k_g p_g + o(1)$$

and that the Lyapunov's condition $\sum_{g=1}^{G} E[|Z_g' u_g/\sqrt{n_g}|^{2+\delta} \cdot (n_g/N)^{1+\delta/2}] \to 0$ holds under Assumption 1, we have that

$$Z'u/\sqrt{N} / \sqrt{\sigma_u^2 \sum_{g=1}^{G} k_g p_g} \Rightarrow N(0,1)$$

by the Lindeberg-Feller CLT, which further implies that

$$\sqrt{N}\left(\hat{\beta}_{pool} - \beta\right)/s_p \Rightarrow N(0,1).$$

where $s_p = \sigma_u \sqrt{\sum_{g=1}^{G} k_g p_g / \left(\sum_{g=1}^{G} \rho_g k_g p_g\right)^2}$.

The fully-interacted estimator $\hat{\beta}_{int} = \beta + (W'P_D W)^{-1} W' P_D u$. It is is consistent because

by the LLN and the rate condition $G^2/N \to 0$,

$$\begin{aligned}
W'P_DW/N &= \sum_{g=1}^{G}(W_g'Z_g/n_g)^2/(Z_g'Z_g/n_g)\cdot n_g/N = \sum_{g=1}^{G}(\rho_g Z_g'Z_g/n_g + v_g'Z_g/n_g)^2/(Z_g'Z_g/n_g)\cdot p_g \\
&= \sum_{g=1}^{G}\rho_g^2 Z_g'Z_g/n_g \cdot p_g + o_p(1) = \sum_{g=1}^{G}\rho_g^2 k_g p_g + o_p(1) \\
W'P_Du/N &= \sum_{g=1}^{G}(W_g'Z_g/n_g)(W_g'u_g/n_g)/(Z_g'Z_g/n_g)\cdot n_g/N = \sum_{g=1}^{G}\rho_g^2 Z_g'u_g/n_g \cdot p_g + o_p(1) = o_p(1) \\
W'P_Du/\sqrt{N} &= \sum_{g=1}^{G}(W_g'Z_g/n_g)(Z_g'u_g/\sqrt{n_g})/(Z_g'Z_g/n_g)\cdot\sqrt{n_g/N} \\
&= \sum_{g=1}^{G}\rho_g(Z_g'u_g/\sqrt{n_g})\cdot\sqrt{n_g/N} + o_p(1)
\end{aligned}$$

Therefore, by the Lindeberg-Feller CLT, we have that

$$\sqrt{N}\left(\hat{\beta}_{int} - \beta\right)/s_{int} \Rightarrow N(0,1).$$

where $s_{int} = \sigma_u/\sqrt{\sum_{g=1}^{G}\rho_g^2 k_g p_g}$.

By the Cauchy-Schwarz inequality, $s_{int} \leq s_p$, since $\left(\sum_{g=1}^{G}\rho_g\sqrt{k_g p_g}\cdot\sqrt{k_g p_g}\right)^2 \leq \left(\sum_{g=1}^{G}\rho_g^2 k_g p_g\right)\cdot\left(\sum_{g=1}^{G}k_g p_g\right)$. The equality holds if and only if $\rho_g = \rho$ for all $g = 1, 2, \ldots, G$. □

Limiting Results of $\hat{\beta}_{pool}$ and $\hat{\beta}_{int}$ in Section 2.2.2 under Causal Effect Heterogeneity:

Proof. Recall that in Section 2.2.2 the causal effect parameter β is replaced by β_g, and only the intercept is included in X (i.e. $X = \ell$, which is a vector of 1); $|\beta_g| \leq \bar{\beta} < \infty$ for all

$g = 1, ..., G$. Then by LLN for independent and not identically distributed sequences,

$$\begin{aligned}
\hat{\beta}_{pool} &= \left(\sum_{g=1}^{G} \left(\rho_g Z_g' Z_g / n_g + Z_g' v_g / n_g \right) p_g \right)^{-1} \left(\sum_{g=1}^{G} \left(\beta_g Z_g' W_g / n_g + \theta_g Z_g' \ell_g / n_g + Z_g' u_g / n_g \right) p_g \right) \\
&= \left(\sum_{g=1}^{G} \rho_g Z_g' Z_g / n_g \cdot p_g + o_p(1) \right)^{-1} \left(\sum_{g=1}^{G} \beta_g \rho_g Z_g' Z_g / n_g \cdot p_g + o_p(1) \right) \\
&= \sum_{g=1}^{G} \frac{\rho_g V_g p_g}{\sum_{g=1}^{G} \rho_g V_g p_g} \beta_g + o_p(1), \\
\hat{\beta}_{int} &= \left(\sum_{g=1}^{G} (W_g' Z_g / n_g)^2 / (Z_g' Z_g / n_g) \cdot p_g \right)^{-1} \sum_{g=1}^{G} (W_g' Z_g / n_g)(Z_g' Y_g / n_g) / (Z_g' Z_g / n_g) \cdot p_g \\
&= \left(\sum_{g=1}^{G} \rho_g^2 Z_g' Z_g / n_g \cdot p_g + o_p(1) \right)^{-1} \left(\sum_{g=1}^{G} \beta_g \rho_g^2 Z_g' Z_g / n_g \cdot p_g + o_p(1) \right) \\
&= \sum_{g=1}^{G} \frac{\rho_g^2 V_g p_g}{\sum_{g=1}^{G} \rho_g^2 V_g p_g} \beta_g + o_p(1).
\end{aligned}$$

with $V_g = V[\tilde{Z}_{ig}]$. □

Limiting Results of $\hat{\beta}_{pool2}$ and $\hat{\beta}_{int2}$ in footnote 3 in Section 2.2.2 :

Proof. Recall that footnote 3 defines $a_g = E[\tilde{Z}_{ig}^2]$ and $b_g = E[\tilde{Z}_{ig}]$. If a single intercept is used in the regression model, the pooled estimator $\hat{\beta}_{pool2} = \left(\tilde{Z}' M_\ell W \right)^{-1} \tilde{Z}' M_\ell Y$. By LLN for independent and not identically distributed sequences,

$$\begin{aligned}
\hat{\beta}_{pool2} &= \left(\tilde{Z}' W / N - \tilde{Z}' \ell / N \cdot W' \ell / N \right)^{-1} \left(\tilde{Z}' Y / N - \tilde{Z}' \ell / N \cdot Y' \ell / N \right) \\
&= \left(\sum_{g=1}^{G} \rho_g p_g \left(a_g - b_g \sum_{g=1}^{G} b_g p_g \right) \right)^{-1} \left(\sum_{g=1}^{G} \beta_g p_g \left(\gamma \left(b_g - \sum_{g=1}^{G} b_g p_g \right) + \rho_g \left(a_g - b_g \sum_{g=1}^{G} b_g p_g \right) \right) \right) + o_p(1).
\end{aligned}$$

The above result holds because

$$\begin{aligned}
\tilde{Z}' W / N &= \sum_g \tilde{Z}_g' (\rho_g \tilde{Z}_g + \gamma \ell_g + v_g) / N = \sum_g \rho_g \tilde{Z}_g' \tilde{Z}_g / N + \gamma \tilde{Z}' \ell / N + o_p(1), \\
\tilde{Z}' \ell / N \cdot W' \ell / N &= \tilde{Z}' \ell / N \cdot \sum_g (\rho_g \tilde{Z}_g + \gamma \ell_g + v_g)' \ell_g / N \\
&= \tilde{Z}' \ell / N \cdot \sum_g \rho_g \tilde{Z}_g' \ell_g / N + \gamma \tilde{Z}' \ell / N + o_p(1),
\end{aligned}$$

implying that the denominator

$$\tilde{Z}'W/N - \tilde{Z}'\ell/N \cdot W'\ell/N = \sum_g \rho_g \tilde{Z}_g'\tilde{Z}_g/N - \sum_g \tilde{Z}_g'\ell_g/N \cdot \sum_g \rho_g \tilde{Z}_g'\ell_g/N + o_p(1)$$

$$= \sum_g \left(\rho_g a_g - \rho_g b_g \cdot \sum_g b_g p_g \right) p_g + o_p(1) \equiv \sum_g DEN_g + o_p(1).$$

Meanwhile,

$$\begin{aligned}
\tilde{Z}'Y/N &= \sum_g \tilde{Z}_g'(\beta_g \rho_g \tilde{Z}_g + \beta_g \gamma \ell_g + \theta \ell_g + \beta_g u_g + v_g)/N \\
&= \sum_g \beta_g \rho_g \tilde{Z}_g'\tilde{Z}_g/N + \gamma \sum_g \beta_g \tilde{Z}_g'\ell_g/N + \theta \tilde{Z}'\ell/N + o_p(1), \\
\tilde{Z}'\ell/N \cdot Y'\ell/N &= \tilde{Z}'\ell/N \cdot \sum_g (\beta_g \rho_g \tilde{Z}_g + \beta_g \gamma \ell_g + \theta \ell_g + \beta_g u_g + v_g)'\ell_g/N \\
&= \tilde{Z}'\ell/N \cdot \sum_g \beta_g \rho_g \tilde{Z}_g'\ell_g/N + \tilde{Z}'\ell/N \cdot \gamma \sum_g \beta_g n_g/N + \theta \tilde{Z}'\ell/N + o_p(1),
\end{aligned}$$

implying that the numerator

$$\tilde{Z}'Y/N - \tilde{Z}'\ell/N \cdot Y'\ell/N = \sum_g \beta_g \rho_g \tilde{Z}_g'\tilde{Z}_g/N - \sum_g \tilde{Z}_g'\ell_g/N \cdot \sum_g \beta_g \rho_g \tilde{Z}_g'\ell_g/N$$

$$+ \gamma \left(\sum_g \beta_g \tilde{Z}_g'\ell_g/N - \tilde{Z}'\ell/N \cdot \sum_g \beta_g n_g/N \right) + o_p(1)$$

$$= \sum_g \beta_g \left(\rho_g a_g - \rho_g b_g \cdot \sum_g b_g p_g + \gamma(b_g - \sum_g b_g p_g) \right) p_g + o_p(1) \equiv \sum_g \beta_g NUM_g + o_p(1).$$

The weight of group g in the probability limit is therefore $NUM_g / \sum_g DEN_g$. Note that the weights sum up to one, or $\sum_g \left(NUM_g / \sum_g DEN_g \right) = 1$ although the first-stage intercept γ enters the weighting formula.

However, the weight $NUM_g / \sum_g DEN_g$ for group g could be negative. This could be seen in a simple example where $\gamma = 0$, $G = 2$, and $p_1 = p_2 = 1/2$. Then the sign of the weights for group $g = 1$ depends on the sign of $NUM_1/(p_1\rho_1) = a_1 - b_1(b_1 + b_2)/2$. Let $a_1 = V_1 + b_1^2$ where V_1 is the variance of $\tilde{Z}_1$, then the sign follows from $V_1 + b_1(b_1 - b_2)/2$. So if b_2 is a lot larger than b_1, the weight for group $g = 1$, or $NUM_1/(DEN_1 + DEN_2)$ could be negative, in which case the weight for group $g = 2$, or $NUM_2/(DEN_1 + DEN_2)$, would be greater than one.

Similarly, if a single intercept is used in the fully-interacted 2SLS regression, the estimator is defined as $\hat{\beta}_{int2} = \left[\left((W\ \ell)' P_{(\tilde{D}\ X)} (W\ \ell)\right)^{-1} (W\ \ell)' P_{(\tilde{D}\ \ell)} Y\right]_1 = \left(W' P_{M_\ell \tilde{D}} W\right)^{-1} W' P_{M_\ell \tilde{D}} Y$. Let $(m_1\ m_2\ ...\ m_G) = W' M_\ell \tilde{D} \left(\tilde{D}' M_\ell \tilde{D}\right)^{-1}$, we have that

$$\hat{\beta}_{int2} = \left(\sum_{g=1}^{G} m_g \tilde{Z}_g' W_g/N - \sum_{g=1}^{G} m_g \tilde{Z}_g' \ell_g/N \cdot W'\ell/N\right)^{-1} \left(\sum_{g=1}^{G} m_g \tilde{Z}_g' Y_g/N - \sum_{g=1}^{G} m_g \tilde{Z}_g' \ell_g/N \cdot Y'\ell/N\right)$$
$$= \left(\sum_{g=1}^{G} \rho_g p_g \left(m_g a_g - b_g \sum_{g=1}^{G} m_g b_g p_g\right)\right)^{-1} \sum_{g=1}^{G} \beta_g p_g \left(\gamma \left(m_g b_g - \sum_{g=1}^{G} m_g b_g p_g\right) + \rho_g \left(m_g a_g - b_g \sum_{g=1}^{G} m_g b_g p_g\right)\right) + o_p(1).$$

□

Proof of Theorem 1

Proof. We only need to prove the theorem for the case where G goes to infinity together with N. In the proof, we will repeatedly use Lemmas shown in Appendix A, as well as the property that the truncated mean monotonically increases with the truncation cutoff since $\frac{\partial}{\partial y} E[X|X > y] = \frac{f_X(y)}{1-F_X(y)}(E[X|X > y] - y) \geq 0$. Further, although Assumption 1 assumes both homoskedasticity and a one-sided first-stage relationship, we will prove this theorem under the more general setting where error term variance is allowed to vary across groups and instruments are allowed to have negative first-stage effects in some groups, denoted by $\mathcal{G}_- = \{g : \rho_g < 0\}$. Assume without loss of generality that $\sigma_{u,v} > 0$. Let $\sigma^2_{g,v} = E[v^2_{ig}]$ with $0 < \underline{\sigma}_v \leq \sigma_{g,v} \leq \bar{\sigma}_v < \infty$ for all $g = 1, ..., G$. Let $G_-{=}|\mathcal{G}_-| = G - G_0 - G_{+,s} - G_{+,w}$.

Since

$$\sqrt{N/G} \cdot E\left[\sum_{g=1}^{G} i_g Z_g' u_g/N_{\alpha_{FS}} \cdot 1\left(N_{\alpha_{FS}} > 0\right)\right] \geq \frac{1}{\sqrt{NG}} E\left[\sum_{g=1}^{G} i_g Z_g' u_g\right],$$

and that u_g and v_g are non-trivially positively correlated, to prove the theorem, it suffices to show that there exists some positive constant a^* such that

$$\frac{1}{G} E\left[\sum_{g=1}^{G} i_g Z_g' v_g/\sqrt{n_g}\right] \geq a^* + o(1).$$

Decomposing the left hand side, we have that

$$\begin{aligned}\frac{1}{G}E\left[\sum_{g=1}^{G} i_g Z_g' v_g/\sqrt{n_g}\right] =& \frac{1}{G}E\left[\sum_{g\in\mathcal{G}_{+,s}} i_g Z_g' v_g/\sqrt{n_g}\right] + \frac{1}{G}E\left[\sum_{g\in\mathcal{G}_0} i_g Z_g' v_g/\sqrt{n_g}\right] \\ &+ \frac{1}{G}E\left[\sum_{g\in\mathcal{G}_{+,w}} i_g Z_g' v_g/\sqrt{n_g}\right] + \frac{1}{G}E\left[\sum_{g\in\mathcal{G}_{-}} i_g Z_g' v_g/\sqrt{n_g}\right] \\ =& A + B + D + E.\end{aligned}$$

All four terms are non-negative as $E\left[\frac{Z_g' v_g}{\sqrt{n_g}}\middle|\frac{Z_g v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2} - \sqrt{n_g} H_{g,2}^2 \rho_g\right] P[t_g > c_g]$ $\geq E\left[\frac{Z_g' v_g}{\sqrt{n_g}}\right] P[t_g > c_g] = 0$. Next, we would like to show that terms B and D are bounded away from zero under the conditions stated in Assumption 1.

Let $\delta_c = \inf_g c_g$ and $\Delta_c = \sup_g c_g$; $\delta_c > 0$ and $\Delta_c < \infty$ for any $0 < \alpha < 1/2$. Let $\mathbf{I}_g = \left\{|H_{g,1}^2 - \sigma_{g,v}^2| \leq \sigma_{g,v}^2/2; |H_{g,2}^2 - k_g| \leq k_g/2\right\}$. Applying Lemma A2, we have that $\frac{1}{G}\sum_g P(\mathbf{I}_g^c) = O(G^2/N^2) = o(1)$. Therefore,

$$\begin{aligned}B &= \frac{1}{G}\sum_{g\in\mathcal{G}_0} E\left[\frac{Z_g' v_g}{\sqrt{n_g}} 1(t_g > c_g)\right] = \frac{1}{G}\sum_{g\in\mathcal{G}_0} E\left[\frac{Z_g' v_g}{\sqrt{n_g}} 1\left(\frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2}\right)\right] \\ &\geq \frac{1}{G}\sum_{g\in\mathcal{G}_0} c_g E\left[H_{g,1} H_{g,2} 1\left(\frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2}\right)\right] \\ &\geq \frac{1}{G}\sum_{g\in\mathcal{G}_0} c_g E\left[H_{g,1} H_{g,2} 1\left(\mathbf{I}_g, \frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2}\right)\right] \\ &\geq \frac{1}{G}\sum_{g\in\mathcal{G}_0} c_g E\left[\frac{1}{2}\sigma_{g,v}\sqrt{k_g} 1\left(\mathbf{I}_g, \frac{Z_g' v_g}{\sqrt{n_g}} > c_g \frac{3}{2}\sigma_{g,v}\sqrt{k_g}\right)\right] \\ &\geq \frac{1}{2}\underline{\sigma}_v\sqrt{\underline{k}}\delta_c \frac{1}{G}\sum_{g\in\mathcal{G}_0}\left(P\left[\frac{Z_g' v_g}{\sqrt{n_g k_g}\sigma_{g,v}} > \frac{3}{2}c_g\right] - P[\mathbf{I}_g^c]\right) \\ &= \frac{1}{2}\underline{\sigma}_v\sqrt{\underline{k}}\delta_c \frac{1}{G}\sum_{g\in\mathcal{G}_0} P\left[\frac{Z_g' v_g}{\sqrt{n_g k_g}\sigma_{g,v}} > \frac{3}{2}c_g\right] + o(1).\end{aligned}$$

For any $g \in \mathcal{G}_0$,

$$
\begin{aligned}
&P\left[\frac{Z_g' v_g}{\sqrt{n_g k_g}\sigma_{g,v}} > \frac{3}{2}c_g\right] = 1 - P\left[\frac{e_g' v_g}{\sqrt{n_g k_g}\sigma_{g,v}} \leq \frac{3}{2}c_g + \frac{e_g' P_{X_g} v_g}{\sqrt{n_g k_g}\sigma_{g,v}}\right] \\
\geq & 1 - P\left[\frac{e_g' v_g}{\sqrt{n_g k_g}\sigma_{g,v}} \leq \frac{3}{2}c_g + \frac{e_g' P_{X_g} v_g}{\sqrt{n_g k_g}\sigma_{g,v}}, \frac{e_g' P_{X_g} v_g}{\sqrt{n_g k_g}\sigma_{g,v}} \leq c_g/2\right] - P\left[\frac{e_g' P_{X_g} v_g}{\sqrt{n_g k_g}\sigma_{g,v}} > c_g/2\right] \\
\geq & 1 - P\left[\frac{e_g' v_g}{\sqrt{n_g k_g}\sigma_{g,v}} \leq 2c_g\right] - P\left[|e_g' P_{X_g} v_g| > \sqrt{n_g k_g}\sigma_{g,v} c_g/2\right].
\end{aligned}
$$

By regularity conditions in Assumption 1, we apply the Berry-Esseen theorem to the first probability term and the Markov and Cauchy-Schwarz inequalities to the second probability term, then we have

$$
B \geq \frac{G_0}{G} \cdot \frac{1}{2}\underline{\sigma}_v \sqrt{\underline{k}}\delta_c \Phi(-2\Delta_c) + o(1).
$$

Similarly, for term D, we have

$$
\begin{aligned}
D &= \frac{1}{G}\sum_{g \in \mathcal{G}_{+,w}} E\left[\frac{Z_g' v_g}{\sqrt{n_g}} \middle| \frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2} - a_g H_{g,2}^2\right] P\left[\frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2} - a_g H_{g,2}^2\right] \\
&\geq \frac{1}{G}\sum_{g \in \mathcal{G}_{+,w}} E\left[\frac{Z_g' v_g}{\sqrt{n_g}} \middle| \frac{Z_g' v_g}{\sqrt{n_g}} > -a_g H_{g,2}^2\right] P\left[\frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2}\right] \\
&\geq \frac{1}{G}\sum_{g \in \mathcal{G}_{+,w}} E\left[\frac{Z_g' v_g}{\sqrt{n_g}} 1\left(\frac{Z_g' v_g}{\sqrt{n_g}} > -a_g H_{g,2}^2\right)\right] P\left[\frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2}\right] \\
&= \frac{1}{G}\sum_{g \in \mathcal{G}_{+,w}} E\left[-\frac{Z_g' v_g}{\sqrt{n_g}} 1\left(\frac{Z_g' v_g}{\sqrt{n_g}} \leq -a_g H_{g,2}^2\right)\right] P\left[\frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2}\right] \\
&\geq \frac{1}{G}\sum_{g \in \mathcal{G}_{+,w}} E\left[a_g H_{g,2}^2 1\left(\frac{Z_g' v_g}{\sqrt{n_g}} \leq -a_g H_{g,2}^2\right) 1(\mathbf{I}_g)\right] P\left[\frac{Z_g' v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2}, \mathbf{I}_g\right] \\
&\geq \frac{\underline{\rho}\underline{k}}{2}\frac{1}{G}\sum_{g \in \mathcal{G}_{+,w}} \left(P\left[\frac{Z_g' v_g}{\sqrt{n_g}} \leq -\frac{\bar{\rho}3k_g}{2}\right] - P(\mathbf{I}_g^c)\right)\left(P\left[\frac{Z_g' v_g}{\sqrt{n_g}} > \frac{c_g 3\sigma_{g,v}\sqrt{k_g}}{2}\right] - P(\mathbf{I}_g^c)\right) \\
&\geq \frac{G_{+,w}}{G} \cdot \frac{\underline{\rho}\underline{k}}{2}\Phi\left(-2\frac{\bar{\rho}\sqrt{\bar{k}}}{\underline{\sigma}_v}\right)\Phi\left(-2\Delta_c\right) + o(1).
\end{aligned}
$$

Under Assumption 1, the term $(G_0 + G_{+,w})/G = 1 - G_{+,s}/G$ is bounded away from zero. Therefore we have that both B and D terms lower bounded by a non-vanishing quantity. Putting together the results stated above, the theorem is proven. □

Proof of Lemma 1

Proof. For the select-and-interact estimator, we have

$$\sqrt{N}(\hat{\beta}_{sel,int}(\delta) - \beta) = \left(\sum_{g=1}^{G} \hat{\rho}_g Z_g' W_g 1(\hat{\mu}_g > \delta)/N\right)^{-1} \sum_{g=1}^{G} \hat{\rho}_g Z_g' u_g 1(\hat{\mu}_g > \delta)/\sqrt{N}.$$

First, consider

$$\begin{aligned}
&\sum_{g=1}^{G} \hat{\rho}_g Z_g' W_g 1(\hat{\mu}_g > \delta)/N = \sum_{g=1}^{G} \left(\rho_g + (Z_g'Z_g)^{-1} Z_g' v_g\right) \left(\rho_g Z_g'Z_g + Z_g' v_g\right) 1(\hat{\mu}_g > \delta)/N \\
=&\frac{1}{N} \sum_{g \in \mathcal{G}_{+,w}} \rho_g^2 Z_g'Z_g 1(\hat{\mu}_g > \delta) + \frac{1}{N} \sum_{g \in \mathcal{G}_{+,s}} \rho_g^2 Z_g'Z_g 1(\hat{\mu}_g > \delta) \\
&+ 2\frac{1}{N} \sum_{g=1}^{G} \rho_g Z_g' v_g 1(\hat{\mu}_g > \delta) + \frac{1}{N} \sum_{g=1}^{G} (Z_g'Z_g)^{-1} (Z_g' v_g)^2 1(\hat{\mu}_g > \delta) \\
=&A_I + A_{II} + B + C.
\end{aligned}$$

First, A_I, B, and C are $o_p(1)$ by Markov's inequliaty and the fact that

$$E\left[|A_I|\right] \leq \frac{1}{N} \sum_{g \in \mathcal{G}_{+,w}} a_g^2 E[Z_g'Z_g/n_g] \leq \bar{k}\bar{\rho} \cdot G_{+,w}/N \to 0,$$

$$E\left[|B|\right] \leq \frac{1}{N} \sum_{g} \rho_g E\left[\left|Z_g' v_g\right|\right] \leq \frac{1}{N} \sum_{g} \rho_g \sqrt{E[(Z_g' v_g)^2]} = \frac{1}{N} \sum_{g} \rho_g \sqrt{\sigma_{g,v}^2 n_g k_g} \leq \sqrt{\bar{\sigma}_v^2 \bar{k} \bar{c} \frac{G}{N}} \to 0.$$

$$E\left[|C|\right] \leq E\left[\frac{1}{N} \sum_{g} (Z_g'Z_g)^{-1} (Z_g' v_g)^2\right] = \frac{1}{N} \sum_{g} \sigma_{g,v}^2 \leq \frac{G}{N} \bar{\sigma}_v^2 \to 0.$$

Then, we prove that $A_{II} = \sum_{g \in \mathcal{G}_{+,s}} \rho_g^2 k_g p_g + o_p(1)$. Let $k'' = \sum_{g \in \mathcal{G}_{+,s}} \rho_g^2 k_g p_g$. Since it is clear $\frac{1}{N} \sum_{g \in \mathcal{G}_{+,s}} \rho_g^2 Z_g'Z_g = k'' + o_p(1)$ as is shown in earlier proofs, it suffices to show that

$$P\left(\left|\frac{1}{N} \sum_{g \in \mathcal{G}_{+,s}} \rho_g^2 Z_g'Z_g (1(\hat{\mu}_g > \delta) - 1)\right| > \epsilon\right) \leq E\left(\left|\frac{1}{N} \sum_{g \in \mathcal{G}_{+,s}} \rho_g^2 Z_g'Z_g (1(\hat{\mu}_g > \delta) - 1)\right|\right)/\epsilon \to 0.$$

Notice that given the δ range in Assumption 2, there exists a small positive constant

$\eta \in (0,1)$ such that $\delta \leq \underline{\rho}\sqrt{\underline{kc}/2}(1-\eta)\sqrt{N/G}$,

$$
\begin{aligned}
&E\left(\left|\frac{1}{N}\sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 Z_g' Z_g (1(\hat{\mu}_g > \delta) - 1)\right|\right) \leq \frac{1}{N}\sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 E\left[Z_g' Z_g 1(\hat{\mu}_g \leq \delta)\right] \\
&\leq \frac{1}{N}\sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 n_g \sqrt{E[(Z_g' Z_g/n_g)^2]}\sqrt{P(\hat{\mu}_g \leq \delta)} \\
&\leq \frac{1}{N}\sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 n_g \sqrt{E[(\tilde{Z}_g' \tilde{Z}_g/n_g)^2]}\sqrt{P\left(\hat{\mu}_g \leq \underline{\rho}\sqrt{\underline{kc}/2}(1-\eta)\sqrt{N/G}\right)} \\
&\leq \frac{1}{N}\sum_{g\in\mathcal{G}_{+,s}} \bar{\rho}^2 \bar{c} N/G \sqrt{E[(\tilde{Z}_{ig}^4)]}\sqrt{P\left(Z_g' Z_g/n_g \leq k_g(1-\eta)\right)} \\
&\to 0.
\end{aligned}
$$

where the convergence result comes from the moment restriction in Assumption 1.3 and a result similar to Lemma A2.

Then we consider

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{g=1}^{G} \hat{\rho}_g Z_g' u_g 1(\hat{\mu}_g > \delta) \\
&= \frac{1}{\sqrt{N}}\sum_g \rho_g Z_g' u_g - \frac{1}{\sqrt{N}}\sum_g \rho_g Z_g' u_g 1(\hat{\mu}_g \leq \delta) + \frac{1}{\sqrt{N}}\sum_g (Z_g' v_g)(Z_g' Z_g)^{-1} Z_g' u_g 1(\hat{\mu}_g > \delta) \\
&= F_1 + F_2 + F_3.
\end{aligned}
$$

Under Assumption 1.3, $F_1/\sqrt{\sigma^2 \cdot k''} \Rightarrow N(0,1)$ by the Lindeberg-Feller CLT. In addition,

F_3 and F_2 are $o_p(1)$ by the Markov's inequality and the fact that

$$
\begin{aligned}
E[|F_3|] &\leq E\left[\frac{1}{\sqrt{N}}\sum_g \left|\frac{Z_g'v_g}{\sqrt{Z_g'Z_g}}\frac{Z_g'u_g}{\sqrt{Z_g'Z_g}}\right|\right] \leq \frac{1}{\sqrt{N}}\sum_g \sqrt{E\left[\left(\frac{Z_g'v_g}{\sqrt{Z_g'Z_g}}\right)^2\right]}\sqrt{E\left[\left(\frac{Z_g'u_g}{\sqrt{Z_g'Z_g}}\right)^2\right]} \\
&\leq \frac{G}{\sqrt{N}}\bar{\sigma}_v\sigma_u \to 0, \\
E[|F_2|] &\leq \frac{1}{\sqrt{N}}\sum_{g\in\mathcal{G}_{+,s}} E[|\rho_g Z_g'u_g 1(\hat{\mu}_g \leq \delta)|] + \frac{1}{\sqrt{N}}\sum_{g\in\mathcal{G}_{+,w}} E[|\rho_g Z_g'u_g 1(\hat{\mu}_g \leq \delta)|] \\
&\leq \frac{1}{\sqrt{N}}\sum_{g\in\mathcal{G}_{+,s}} \rho_g\sqrt{E\left[(Z_g'u_g)^2\right]}\sqrt{P(\hat{\mu}_g \leq \delta)} + \frac{1}{\sqrt{N}}\sum_{g\in\mathcal{G}_{+,w}} \rho_g E[|Z_g'u_g|] \\
&\leq \frac{1}{\sqrt{N}}\sum_{g\in\mathcal{G}_{+,s}} \bar{\rho}n_g\sqrt{E\left[(Z_g'u_g/\sqrt{n_g})^2\right]}\sqrt{P(\hat{\mu}_g \leq \delta)} + \frac{1}{\sqrt{N}}\sum_{g\in\mathcal{G}_{+,w}} \bar{\rho}\sqrt{E[(Z_g'u_g/\sqrt{n_g})^2]} \\
&= O(G/\sqrt{N}\cdot N/G\cdot 1/(N/G)) + O(G/\sqrt{N}) = o(1).
\end{aligned}
$$

Note that the convergence result follows from the moment restrictions in Assumption 1 and the fact that $P(\hat{\mu}_g \leq \delta) \leq P\left(Z_g'Z_g/n_g \leq k_g(1-\eta)\right) \lesssim 1/n_g^2$ given the δ range and Lemma A2.

Combining results and applying Slutsky's Theorem, we obtain

$$\sqrt{N}(\hat{\beta}_{sel,int}(\delta) - \beta)/\sqrt{\sigma_u^2/k''} \Rightarrow N(0,1).$$

Notice that $s_{sel,int} = \sqrt{\sigma_u^2/k''}$, the first part of the lemma is proven.

For the second part of the lemma, first we consider the asymptotic property of $\hat{\beta}^a(\delta)$, the split-sample estimator with sample a. Notice that we have

$$
\begin{aligned}
&\sqrt{N^a}(\hat{\beta}^a(\delta) - \beta) \\
&= \left(\frac{1}{N^a}\sum_g (W_g^b)'Z_g^b(Z_g^{b'}Z_g^b)^{-1}(Z_g^a)'W_g^a 1(\hat{\mu}_g^b > \delta)\right)^{-1} \frac{1}{\sqrt{N^a}}\sum_g (W_g^b)'Z_g^b((Z_g^b)'Z_g^b)^{-1}(Z_g^a)'u_g^a 1(\hat{\mu}_g^b > \delta).
\end{aligned}
$$

Note that the denominator follows that

$$
\begin{aligned}
&\frac{1}{N^a}\sum_g (W_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1} (Z_g^a)' W_g^a 1(\hat{\mu}_g^b > \delta) \\
=&\frac{1}{N^a}\sum_g \left(\rho_g + ((Z_g^b)' Z_g^b)^{-1}(Z_g^b)' v_g^b\right)\left(\rho_g (Z_g^a)' Z_g^a + (Z_g^a)' v_g^a\right) 1(\hat{\mu}_g^b > \delta) \\
=&\frac{1}{N^a}\sum_g \rho_g^2 (Z_g^a)' Z_g^a 1(\hat{\mu}_g^b > \delta) + \frac{1}{N^a}\sum_g \rho_g (Z_g^a)' v_g^a 1(\hat{\mu}_g^b > \delta) \\
&+ \frac{1}{N^a}\sum_g \rho_g (Z_g^a)' Z_g^a ((Z_g^b)' Z_g^b)^{-1}(Z_g^b)' v_g^b 1(\hat{\mu}_g^b > \delta) + \frac{1}{N^a}\sum_g ((Z_g^b)' Z_g^b)^{-1}(Z_g^b)' v_g^b (Z_g^a)' v_g^a 1(\hat{\mu}_g^b > \delta) \\
=&\frac{1}{N^a}\sum_g \rho_g^2 Z_g^{a'} Z_g^a 1(\hat{\mu}_g^b > \delta) + o_p(1), \\
=&k'' + o_p(1).
\end{aligned}
$$

The numerator

$$
\begin{aligned}
&\sum_g (W_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1} (Z_g^a)' u_g^a 1(\hat{\mu}_g^b > \delta)/\sqrt{N^a} \\
=&\frac{1}{\sqrt{N^a}}\sum_g \rho_g (Z_g^a)' u_g^a - \frac{1}{\sqrt{N^a}}\sum_g \rho_g (Z_g^a)' u_g^a 1(\hat{\mu}_g^b \le \delta) + \frac{1}{\sqrt{N^a}}((Z_g^b)' Z_g^b)^{-1}(Z_g^b)' v_g^b (Z_g^b)' u_g^a 1(\hat{\mu}_g^b > \delta) \\
=&\frac{1}{\sqrt{N^a}}\sum_g \rho_g (Z_g^a)' u_g^a + o_p(1).
\end{aligned}
$$

Similar derivations also hold for the split-sample estimator in sample b. Putting them together, we obtain

$$
\begin{aligned}
\sqrt{N}(\hat{\beta}_{sssel,int}(\delta) - \beta) &= \frac{1}{2}\sqrt{N/N^a}\sqrt{N^a}(\hat{\beta}^a(\delta) - \beta) + \frac{1}{2}\sqrt{N/N^b}\sqrt{N^b}(\hat{\beta}^b(\delta) - \beta) \\
&= \frac{1}{\sqrt{2}}(h^a + h^b)/k'' + o_p(1)
\end{aligned}
$$

where $h^a = \frac{1}{\sqrt{N^a}}\sum_g \rho_g Z_g^{a'} u_g^a$ and $h^b = \frac{1}{\sqrt{N^b}}\sum_g \rho_g^2 Z_g^{b'} u_g^b$. Since h^a and h^b are independent, we have that $(h^a/\sqrt{\sigma_u^2 k''}, h^b/\sqrt{\sigma_u^2 k''})'$ converges jointly to $N((0,0)', I_2)$. Therefore,

$$
\sqrt{N}(\hat{\beta}_{sssel,int}(\delta) - \beta)/s_{sel,int} \Rightarrow N(0,1).
$$

□

Proof for Theorem 2

Proof. We first state Lemma A.1 of Donald and Newey (2001) in the following Lemma A6. Since instead of choosing K as in Donald and Newey (2001), we are choosing the cutoff value δ, the decomposition in the following Lemma depends on δ.

Lemma A6 (Donald and Newey (2001) Lemma A.1)**.** *Suppose the estimator examined has the form* $\sqrt{N}(\hat{\beta} - \beta_0) = \hat{H}^{-1}\hat{h}$. *If there is a decomposition* $\hat{h} = h + T^h + Z^h$ *and* $\hat{H} = H + T^H + Z^H$ *and*

$$(h + T^h)^2 - 2h^2 H^{-1} T^H = \hat{A}(\delta) + Z^A(\delta)$$

such that

1) $h = O_p(1)$, $H = O_p(1)$, 2) $\sup_\delta S(\delta) = o_p(1)$, 3) $\sup_\delta T^h = o_p(1)$,

4) $\sup_\delta \left((T^H)^2/S(\delta)\right) = o_p(1)$, 5) $\sup_\delta \left((T^H)(T^h)/S(\delta)\right) = o_p(1)$, 6) $\sup_\delta \left(Z^H/S(\delta)\right) = o_p(1)$,

7) $\sup_\delta \dfrac{E[\hat{A}(\delta)|\tilde{Z}, \tilde{X}] - \sigma_u^2 H - S(\delta)H^2}{S(\delta)} = o_p(1)$.

8) $\sup_\delta \left(Z^A(\delta)/S(\delta)\right) = o_p(1)$

then

$$N(\hat{\beta} - \beta_0)^2 = \hat{Q}(\delta) + \hat{r}(\delta)$$
$$E[\hat{Q}(\delta)|\tilde{Z}, \tilde{X}] = \sigma_u^2 H^{-1} + S(\delta) + T(\delta)$$
$$\sup_\delta \left\{(\hat{r}(\delta) + T(\delta))/S(\delta)\right\} = o_p(1) \text{ as } G, N \to \infty.$$

Note that the asymptotic MSE decomposition in Lemma A6 refers to the mean squared error of the estimator of interest conditional on exogenous variables. Following the proofs of Donald and Newey (2001), we omit the conditioning from the expectation for the rest of the proof for notational simplicity.

First, we prove the first statement of the theorem for the select-and-interact estimator $\hat{\beta}_{sel,int}$. Note that the estimator has the form

$$\sqrt{N}(\hat{\beta}_{sel,int}(\delta) - \beta) = \hat{H}_\delta^{-1}\hat{h}_\delta$$

where $\hat{h}_\delta = \frac{W'P_\delta u}{\sqrt{N}}$ and $\hat{H}_\delta = \frac{W'P_\delta W}{N}$ and P_δ is a block diagonal matrix consisting of matrices $Z_g(Z_g'Z_g)^{-1}Z_g'1(\hat{\mu}_g > \delta)$ on its diagonals. Let $f = [\rho_1 Z_1' \ \rho_2 Z_2' \ ... \ \rho_G Z_G']'$, $\hat{h}_\delta = h + T_1^h + T_2^h$, and $\hat{H}_\delta = H + T_1^H + T_2^H + T_3^H + Z^H$ with

$$h = f'u/\sqrt{N} = \sum_g \rho_g Z_g'u_g/\sqrt{N}; \quad H = \sum_{g\in\mathcal{G}_{+,s}} \rho_g^2 Z_g'Z_g/N = \sum_{g\in\mathcal{G}_{+,s}} \mu_g^2/N;$$
$$T_1^h(\delta) = -f'(I-P_\delta)u/\sqrt{N} = -\sum_g \rho_g 1(\hat{\mu}_g < \delta)Z_g'u_g/\sqrt{N};$$
$$T_2^h(\delta) = v'P_\delta u/\sqrt{N} = \sum_g v_g'Z_g(Z_g'Z_g)^{-1}Z_g'u_g 1(\hat{\mu}_g \geq \delta)/\sqrt{N}; \quad T_1^H = f'f/N - H = \sum_{g\in\mathcal{G}_{+,w}} \mu_g^2/N;$$
$$T_2^H(\delta) = -f'(I-P_\delta)f/N; \quad T_3^H = (v'f + f'v)/N; \quad Z^H(\delta) = (v'P_\delta v - v'(I-P_\delta)f - f'(I-P_\delta)v)/N.$$

Now conforming to the notations in Lemma A6, let $Z^A(\delta) = 0$ hence $\hat{A}_{sel,int}(\delta) = (h + T_1^h(\delta) + T_2^h(\delta))^2 - 2h^2H^{-1}(T_1^H + T_2^H(\delta) + T_3^H)$.

Denote $t_{g,\delta} = (\delta - \mu_g)/\sigma_v$. Let $\Phi_{g,\delta} = \Phi(t_{g,\delta})$ and $\phi_{g,\delta} = \phi(t_{g,\delta})$. By the normality assumption of error terms, we are able to simplify the following expectations:

$$E[1(\hat{\mu}_g > \delta)] = 1 - \Phi_{g,\delta}$$
$$E[1(\hat{\mu}_g > \delta)Z_g'v_g] = \sigma_v\phi_{g,\delta}\sqrt{Z_g'Z_g}$$
$$E[1(\hat{\mu}_g > \delta)\left(Z_g'v_g\right)^2] = \sigma_v^2\left(1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta}\right)Z_g'Z_g$$
$$E[1(\hat{\mu}_g > \delta)\left(Z_g'v_g\right)^3] = \sigma_v^3\phi_{g,\delta}\left(t_{g,\delta}^2 + 2\right)\left(Z_g'Z_g\right)^{3/2}$$
$$E[1(\hat{\mu}_g > \delta)\left(Z_g'v_g\right)^4] = \sigma_v^4\phi_{g,\delta}\left(t_{g,\delta}^3 + 3t_{g,\delta}\right)\left(Z_g'Z_g\right)^2 + 3(1-\Phi_{g,\delta})(Z_g'Z_g)^2$$
$$E[1(\hat{\mu}_g > \delta)Z_g'u_g] = \frac{\sigma_{uv}}{\sigma_v}\phi_{g,\delta}\sqrt{Z_g'Z_g}$$
$$E[1(\hat{\mu}_g > \delta)(Z_g'u_g)(Z_g'v_g)] = \frac{\sigma_{uv}}{\sigma_v^2}E[1(\hat{\mu}_g > \delta)(Z_g'v_g)^2] = \sigma_{uv}\left(1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta}\right)Z_g'Z_g$$
$$E[1(\hat{\mu}_g > \delta)(Z_g'u_g)^2] = \frac{\sigma_{uv}^2}{\sigma_v^4}E[1(\hat{\mu}_g > \delta)(Z_g'v_g)^2] + (\sigma_u^2 - \frac{\sigma_{uv}^2}{\sigma_v^2})E[1(\hat{\mu}_g > \delta)]Z_g'Z_g$$
$$= \left(\sigma_u^2(1-\Phi_{g,\delta}) + \frac{\sigma_{uv}^2}{\sigma_v^2}t_{g,\delta}\phi_{g,\delta}\right)Z_g'Z_g.$$

Asymptotic MSE for the Select-and-Interact Estimator

Given Lemma A6, we know that to prove that the first part of the theorem, we just need

to prove that under Assumption 2,

$$1)\ h = O_p(1),\quad H = O_p(1),\quad 2)\ \sup_{\delta\in\Delta} S_{sel,int}(\delta) = o_p(1),\quad 3)\ \sup_{\delta\in\Delta} T_1^h + T_2^h = o_p(1),$$

$$4)\ \sup_{\delta\in\Delta}\Big((T_1^H + T_2^H + T_3^H)^2/S_{sel,int}(\delta)\Big) = o_p(1),$$

$$5)\ \sup_{\delta\in\Delta}\Big((T_1^H + T_2^H + T_3^H)(T_1^h + T_2^h)/S_{sel,int}(\delta)\Big) = o_p(1),\quad 6)\ \sup_{\delta\in\Delta}\Big(Z^H/S_{sel,int}(\delta)\Big) = o_p(1),$$

$$7)\ \sup_{\delta\in\Delta}\frac{E[\hat{A}_{sel,int}(\delta)] - \sigma_u^2 H - H^2 S_{sel,int}(\delta)}{S_{sel,int}(\delta)} = o_p(1).$$

To prove all seven statements above hold, we take the following steps: (1) we decompose $E[\hat{A}_{sel,int}(\delta)]$; 2) we show that $S_{sel,int}(\delta)$ defined in the theorem is the right higher-order leading term such that the seventh statement above holds; (3) we prove the remaining six statements.

Step 1: Decomposition of $E[\hat{A}_{sel,int}(\delta)]$

Note that

$$\begin{aligned}
E[\hat{A}_{sel,int}(\delta)] =& E[(h + T_1^h + T_2^h)^2] - 2E[h^2 H^{-1}(T_1^H + T_2^H + T_3^H)] \\
=& \sigma_u^2 H + \big(E[(h + T_1^h)^2] - \sigma_u^2 H\big) + 2E[(h + T_1^h)T_2^h] + E[(T_2^h)^2] - 2E[h^2 H^{-1} T_1^H] \\
& - 2E[h^2 H^{-1} T_2^H] - -2E[h^2 H^{-1} T_3^H] \\
=& \sigma_u^2 H + \Delta_1(\delta) + \Delta_2(\delta) + \Delta_3(\delta) + \Delta_4 + \Delta_5(\delta) + \Delta_6.
\end{aligned}$$

For the $\Delta_1(\delta)$ term,

$$\begin{aligned}
\Delta_1(\delta) =& E[(f'P_\delta u)^2/N] - \sigma_u^2 H \\
=& E\left[\left(\sum_g \rho_g Z_g' u_g 1(\hat{\mu}_g > \delta)\right)^2\right]/N - \sigma_u^2 \sum_{g\in\mathcal{G}_{+,s}} \mu_g^2/N \\
=& \sum_g E[\rho_g^2 (Z_g' u_g)^2 1(\hat{\mu}_g > \delta)]/N + \left(\sum_g E[\rho_g Z_g' u_g 1(\hat{\mu}_g > \delta)]\right)^2/N \\
& - \sum_g E[\rho_g Z_g' u_g 1(\hat{\mu}_g > \delta)]^2 - \sigma_u^2 \sum_{g\in\mathcal{G}_{+,s}} \mu_g^2/N \\
=& \sigma_u^2 \sum_{g\in\mathcal{G}_{+,w}} \mu_g^2/N + \frac{\sigma_{uv}^2}{\sigma_v^2}\left(\sum_g \mu_g \phi_{g,\delta}\right)^2/N \\
& - \sigma_u^2 \sum_g \mu_g^2 \Phi_{g,\delta}/N + \frac{\sigma_{uv}^2}{\sigma_v^2}\sum_g \mu_g^2 t_{g,\delta}\phi_{g,\delta}/N - \frac{\sigma_{uv}^2}{\sigma_v^2}\sum_g \mu_g^2 \phi_{g,\delta}^2/N.
\end{aligned}$$

Now use results in Lemma A4, we have

$$\sup_{\delta\in\Delta} \Delta_1(\delta) = \sup_{\delta\in\Delta} \frac{\sigma_{uv}^2}{\sigma_v^2}\left(\sum_g \mu_g \phi_{g,\delta}\right)^2/N + O_p(G/N).$$

For the $\Delta_2(\delta)$ terms, we have

$$
\begin{aligned}
\Delta_2(\delta) =& 2E[(f'P_\delta u)(v'P_\delta u)/N] \\
=& 2E\left[\sum_g \rho_g Z_g' u_g 1(\hat{\mu}_g \geq \delta) \sum_g v_g' Z_g (Z_g'Z_g)^{-1} Z_g' u_g 1(\hat{\mu}_g \geq \delta)/N\right] \\
=& 2\sum_g E\left[\rho_g (Z_g'Z_g)^{-1}(Z_g'u_g)^2 (Z_g'v_g) 1(\hat{\mu}_g \geq \delta)\right]/N \\
&+ 2\sum_g E[\rho_g Z_g' u_g 1(\hat{\mu}_g \geq \delta)] \sum_g E[v_g' Z_g (Z_g'Z_g)^{-1} Z_g' u_g 1(\hat{\mu}_g \geq \delta)]/N \\
&- 2\sum_g E[\rho_g Z_g' u_g 1(\hat{\mu}_g \geq \delta)] E[v_g' Z_g (Z_g'Z_g)^{-1} Z_g' u_g 1(\hat{\mu}_g \geq \delta)]/N \\
=& 2\frac{\sigma_{uv}^2}{\sigma_v} \sum_g \mu_g \phi_{g,\delta} \left(t_{g,\delta}^2 + 2\right)/N + 2\sigma_v(\sigma_u^2 - \frac{\sigma_{uv}^2}{\sigma_v^2}) \sum_g \mu_g \phi_{g,\delta}/N \\
&+ 2\frac{\sigma_{uv}^2}{\sigma_v} \sum_g \mu_g \phi_{g,\delta} \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta})/N - 2\frac{\sigma_{uv}^2}{\sigma_v} \sum_g \mu_g \phi_{g,\delta}(1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta})/N \\
=& 2\frac{\sigma_{uv}^2}{\sigma_v} \sum_g \mu_g \phi_{g,\delta} t_{g,\delta}^2/N + 2\sigma_v\sigma_u^2 \sum_g \mu_g \phi_{g,\delta}/N \\
&+ 2\frac{\sigma_{uv}^2}{\sigma_v} \sum_g \mu_g \phi_{g,\delta} \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta})/N + 2\frac{\sigma_{uv}^2}{\sigma_v} \sum_g \mu_g \phi_{g,\delta}(\Phi_{g,\delta} - t_{g,\delta}\phi_{g,\delta})/N.
\end{aligned}
$$

Note that $\Phi(x) - x\phi(x)$ is monotonically increasing and therefore $0 \leq \Phi(x) - x\phi(x) \leq 1$, then by results in Lemma A4, we have

$$
\sup_{\delta \in \Delta} \Delta_2(\delta) = \sup_{\delta \in \Delta} 2\frac{\sigma_{uv}^2}{\sigma_v} \left(\sum_g \mu_g \phi_{g,\delta}\right) \left(\sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta})\right)/N + O_p(G/N),
$$

where the last equality holds as $\Phi(x) - x\phi(x)$ is monotonically increasing and therefore $0 \leq \Phi(x) - x\phi(x) \leq 1$ and by convergence results derived in Lemma A4.

For the $\Delta_3(\delta)$ term, we have

$$\begin{aligned}\Delta_3(\delta) =& E[(v'P_\delta u)^2/N] = E\left[\left(\sum_g v_g' Z_g (Z_g'Z_g)^{-1} Z_g' u_g 1(\hat{\mu}_g > \delta)\right)^2 /N\right] \\ =& \sum_g E[(v_g'Z_g)^2 (Z_g'Z_g)^{-2} (Z_g'u_g)^2 1(\hat{\mu}_g > \delta)]/N + \left[\sum_g E[v_g'Z_g(Z_g'Z_g)^{-1}Z_g'u_g 1(\hat{\mu}_g > \delta)]\right]^2 /N \\ & - \sum_g E\left[v_g'Z_g(Z_g'Z_g)^{-1}Z_g'u_g 1(\hat{\mu}_g > \delta)\right]^2 /N \\ =& \sigma_{uv}^2 \sum_g (t_{g,\delta}^3 + 3t_{g,\delta})\phi_{g,\delta}/N + \sigma_{uv}^2 \sum_g 3(1-\Phi_{g,\delta})/N + (\sigma_u^2\sigma_v^2 - \sigma_{uv}^2)\sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta})/N \\ & + \sigma_{uv}^2 \left(\sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta})\right)^2 /N - \sigma_{uv}^2 \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta})^2 /N\end{aligned}$$

Then, since $0 \leq 1 - \Phi(x) + x\phi(x) \leq 1$ and the results in Lemma A4,

$$\sup_{\delta \in \Delta} \Delta_3(\delta) = \sup_{\delta \in \Delta} \sigma_{uv}^2 \left(\sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta})\right)^2 /N + O_p(G/N).$$

For the Δ_4 term, notice that

$$\Delta_4 = -2E[h^2 H^{-1} T_1^H] = -2E[h^2]H^{-1}T_1^H = -2\sigma_u^2(H + T_1^H)T_1^H/H = O_p(G/N)$$

as $H \xrightarrow{p} k''$ and $T_1^H = O_p(G/N)$ as is shown in the proof of Lemma 1.

For the $\Delta_5(\delta)$ term,

$$\begin{aligned}0 \leq \Delta_5(\delta) =& 2E\left[\left(\sum_g \rho_g Z_g' u_g\right)^2 \sum_g \rho_g^2 Z_g' Z_g 1(\hat{\mu}_g < \delta)\right] /(N^2 H) \\ =& 2\sum_g E\left[\rho_g^4 (Z_g'u_g)^2 Z_g'Z_g 1(\hat{\mu}_g < \delta)\right] /(N^2H) \\ & + 2\sum_g E[\rho_g^2 (Z_g'u_g)^2] \sum_g E[\rho_g^2 Z_g'Z_g 1(\hat{\mu}_g < \delta)]/(N^2H) \\ & - 2\sum_g E[\rho_g^2 (Z_g'u_g)^2] E[\rho_g^2 Z_g'Z_g 1(\hat{\mu}_g < \delta)]/(N^2H) \\ =& 2\sum_g \mu_g^4 \Big(\sigma_u^2 \Phi_{g,\delta} - \frac{\sigma_{uv}^2}{\sigma_v^2} t_{g,\delta}\phi_{g,\delta}\Big)/(N^2H) + 2\sigma_u^2 \sum_g \mu_g^2 \Phi_{g,\delta}(H + T_1^H)/(NH) \\ & - 2\sigma_u^2 \sum_g \mu_g^4 \Phi_{g,\delta}/(N^2H).\end{aligned}$$

Apply the results derived in Lemma A4, we have $\sup_{\delta\in\Delta}\Delta_5(\delta) = O_p(G/N)$.

The last term $\Delta_6 = -2E[h^2H^{-1}T_3^H] = -2E[(f'u)^2(v'f + f'v)]/H = 0$ by symmetry of normal distributions.

Step 2: determine $S_{sel,int}(\delta)$

Notice that adding up Δ_1 to Δ_6, we get

$$\begin{aligned}E[\hat{A}_{sel,int}(\delta)] =&\sigma_{uv}^2\left(\sum_g(1-\Phi_{g,\delta}+t_{g,\delta}\phi_{g,\delta})\right)^2/N + 2\frac{\sigma_{uv}^2}{\sigma_v}\left(\sum_g\mu_g\phi_{g,\delta}\right)\left(\sum_g(1-\Phi_{g,\delta}+t_{g,\delta}\phi_{g,\delta})\right)/N\\ &+\frac{\sigma_{uv}^2}{\sigma_v^2}\left(\sum_g\mu_g\phi_{g,\delta}\right)^2/N + O_p(G/N)\\ =&\left(\sigma_{uv}\sum_g(1-\Phi_{g,\delta}+t_{g,\delta}\phi_{g,\delta}) + \frac{\sigma_{uv}}{\sigma_v}\sum_g\mu_g\phi_{g,\delta}\right)^2/N + O_p(G/N).\end{aligned}$$

Notice that $0 \le 1-\Phi_{g,\delta}+t_{g,\delta}\phi_{g,\delta} \le 1$ and $\sum_g\mu_g\phi_{g,\delta} = O_p(G)$ following convergence results derived in Lemma A4. Set $S_{sel,int}(\delta)H^2 = \left(\sigma_{uv}\sum_g(1-\Phi_{g,\delta}+t_{g,\delta}\phi_{g,\delta}) + \frac{\sigma_{uv}}{\sigma_v}\sum_g\mu_g\phi_{g,\delta}\right)^2/N$, we know that $\sup_{\delta\in\Delta}S_{sel,int}(\delta) = O_p(G^2/N)$. Further, since $\sum_{g\in\mathcal{G}_{+,s}}(\Phi_{g,\delta}-t_{g,\delta}\phi_{g,\delta}) = O_p(G^3/N^2)$ following convergence results derived in Lemma A4, we also know that $\inf_{\delta\in\Delta}S_{sel,int}(\delta)H^2 \ge \sigma_{uv}^2G_{+,s}^2/N + o_p(G^2/N) = \sigma_{uv}^2b^2G^2/N + o_p(G^2/N)$ for some strictly positive b following Assumption 1. Therefore, any term of order $o_p(G^2/N)$ is dominated by $S_{sel,int}(\delta)$ uniformly over $\delta\in\Delta$.

Step 3: Proof of the corresponding statement (1) - (6) in Lemma A6

For statement (1), both $h = O_p(1)$ and $H = O_p(1)$ have been shown in the proof of Lemma 1.

For statement (2), $\sup_{\delta\in\Delta}S_{sel,int}(\delta) = o_p(1)$ as $G^2/N \to 0$.

For statement (3), note that $T_1^h = o_p(1)$ follows from the Markov inequality and the fact that

$$\begin{aligned}E[(T_1^h)^2 \le&\sqrt{E[(T_1^h)^2]} = \sigma_u^2\sum_g\mu_g^2\Phi_{g,\delta}/N - \frac{\sigma_{uv}^2}{\sigma_v^2}\sum_g\mu_g^2t_{g,\delta}\phi_{g,\delta}/N\\ &-\frac{\sigma_{uv}^2}{\sigma_v^2}\sum_g\mu_g^2\phi_{g,\delta}^2/N + \frac{\sigma_{uv}^2}{\sigma_v^2}\Big[\sum_g\mu_g\phi_{g,\delta}\Big]^2/N\end{aligned}$$

Applying Lemma A4, we have $\sup_{\delta\in\Delta} T_1^h(\delta) = O_p(G^2/N) = o_p(1)$. Similarly $\sup_{\delta\in\Delta} T_2^h(\delta) = o_p(1)$ follows from the fact that $\sup_{\delta\in\Delta} \Delta_3(\delta) = O_p(G^2/N) = o_p(1)$.

For statement (4), note that $T_1^H = O_p(G/N)$ by following Assumption 1. $\sup_{\delta\in\Delta} T_2^H(\delta) = O_p(G/N)$ by Markov inequality and the fact that $\sup_{\delta\in\Delta} E[|T_2^H|] = \sup_{\delta\in\Delta} \frac{1}{N}\sum_g \mu_g^2\Phi_{g,\delta} = O_p(G/N)$. $T_3^H = O_p(1/\sqrt{N})$ by the central limit theorem. Since each of G^2/N^2, N^{-1} and $G/N/\sqrt{N}$ is $o_p(G^2/N)$, $\sup_{\delta\in\Delta}\left((T_1^H + T_2^H + T_3^H)^2/S_{sel,int}(\delta)\right) = o_p(1)$.

For statement (5), note that $\sup_{\delta\in\Delta} T_1^h(\delta) + T_2^h(\delta) = O_p(G^2/N)$ and $T_1^H + \sup_{\delta\in\Delta} T_2^H + T_3^H = O_p(G/N + 1/\sqrt{N})$. Therefore, their product is $o_p(G^2/N)$, and hence $\sup_{\delta\in\Delta}\left((T_1^H + T_2^H + T_3^H)(T_1^h + T_2^h)/S_{sel,int}(\delta)\right) = o_p(1)$.

Lastly, for statement (6), note that $Z^H(\delta) = \frac{v'P_\delta v - v'(I-P_\delta)f - f'(I-P_\delta)v}{N}$. The first term is $v'P_\delta v/N = tr(P_\delta E[vv'|X])/N = O_p(G/N)$ uniformly over $\delta\in\Delta$. The second and third terms are $O_p(G/N)$ uniformly over $\delta\in\Delta$ by the Markov inequality. Combine the fact that $E\left[v'(I-P_\delta)f/N\right] \le \sqrt{E\left[(v'(I-P_\delta)f/N)^2\right]}$, and that

$$
\begin{aligned}
&E\left[(v'(I-P_\delta)f/N)^2\right] = E\left[\left(\sum_g \rho_g Z_g' v_g 1(\hat{\mu}_g < \delta)\right)^2\right]/N^2 \\
=&\sum_g \rho_g^2 E[(Z_g'v_g)^2 1(\hat{\mu}_g<\delta)]/N^2 + \left(\mathbb{E}\left[\sum_g \rho_g Z_g' v_g 1(\hat{\mu}_g<\delta)\right]\right)^2/N^2 \\
&-\sum_g \left(E\left[\rho_g Z_g' v_g 1(\hat{\mu}_g<\delta))^2\right]\right)^2/N^2 \\
=&\sum_g \mu_g^2\sigma_v^2(\Phi_{g,\delta} - t_{g,\delta}\phi_{g,\delta})/N^2 + \left(\sum_g \mu_g\sigma_v\phi_{g,\delta}\right)^2/N^2 - \sum_g \left(\mu_g\sigma_v\phi_{g,\delta}\right)^2/N^2.
\end{aligned}
$$

Applying results in Lemma A4, we have $\sup_{\delta\in\Delta}\left(Z_H(\delta)/S_{sel,int}(\delta)\right) = o_p(1)$.

Following the three steps, the first part of the theorem is proven.

Asymptotic MSE for the Repeated Split-Sample Select-and-Interact Estimator

Now we prove the second part of the theorem. To facilitate the proof, we first provide the result for the MSE decomposition for the split-sample 2SLS estimator using half of the sample in the following lemma. Let $\mu_g^a = \rho_g\sqrt{(Z_g^a)'Z_g^a}$, $\hat{\mu}_g^a = ((Z_g^a)'Z_g^a)^{-1/2}(Z_g^a)'W_g^a$, $\Phi_{g,\delta}^a = \Phi(\frac{\delta-\mu_g^a}{\sigma_v})$, $\phi_{g,\delta}^a = \phi(\frac{\delta-\mu_g^a}{\sigma_v})$, $t_{g,\delta}^a = \frac{\delta-\mu_g^a}{\sigma_v}$, and define similar expressions for subsample b.

Lemma A7. *Under Assumptions stated in Thereom 2, the asymptotic MSE of $\hat{\beta}^a$ follows the decomposition*

$$\begin{aligned} N^a(\hat{\beta}^a(\delta)-\beta)^2 &= \hat{Q}^a(\delta)+\hat{r}^a(\delta), \\ E[\hat{Q}^a(\delta)|\tilde{Z},X] &= \sigma_u^2(H^a)^{-1}+S^a(\delta)+T^a(\delta), \\ \sup_{\delta\in\Delta}\Big((\hat{r}^a(\delta)+T^a(\delta))/S^a(\delta)\Big) &= o_p(1), \end{aligned}$$

with $H^a=\sum_g \rho_g^2 Z_g^{a'}Z_g^a/N_a=\sum_g(\mu_g^a)^2/N_a$ *and* $(H^a)^2S^a(\delta)=\sigma_u^2\sigma_v^2\sum_g\left(1-\Phi_{g,\delta}^b+t_{g,\delta}^b\phi_{g,\delta}^b\right)/N^a+\sigma_u^2\sum_g(\mu_g^a)^2\Phi_{g,\delta}^b/N^a$.

Proof. Similar to the proof in the first part of the theorem, we first specify the terms in the $E[\hat{Q}^a(\delta)|\tilde{Z},X])$, and then verify the conditions of Lemma A6 hold with the $S^a(\delta)$ defined in the lemma.

First, notice that $\sqrt{N^a}(\hat{\beta}^a(\delta)-\beta)=(\hat{H}_\delta^a)^{-1}\hat{h}_\delta^a$, where $\hat{h}_\delta^a=\sum_g W_g^{b'}Z_g^b(Z_g^{b'}Z_g^b)^{-1}Z_g^{a'}u_g^a1(\hat{\mu}_g^b>\delta)/\sqrt{N^a}$ and $\hat{H}_\delta^a=\sum_g W_g^{b'}Z_g^b(Z_g^{b'}Z_g^b)^{-1}Z_g^{a'}W_g^a1(\hat{\mu}_g^b>\delta)/N^a$. Let $\hat{h}_\delta^a=h^a+T_{1h}^a(\delta)+T_{2h}^a(\delta)$ and $\hat{H}_\delta^a=H^a+T_{1H}^a(\delta)+T_{2H}^a(\delta)+Z_H^a(\delta)$, where

$$\begin{aligned} h^a &= \sum_g \rho_g Z_g^{a'}u_g^a/\sqrt{N^a};\quad H^a=\sum_g\rho_g^2Z_g^{a'}Z_g^a/N^a;\quad T_{1h}^a(\delta)=-\sum_g\rho_g1(\hat{\mu}_g^b<\delta)Z_g^{a'}u_g^a/\sqrt{N^a}; \\ T_{2h}^a(\delta) &= \sum_g v_g^{b'}Z_g^b(Z_g^{b'}Z_g^b)^{-1}Z_g^{a'}u_g^a1(\hat{\mu}_g^b\geq\delta)/\sqrt{N^a};\quad T_{1H}^a(\delta)=-\sum_g\rho_g^2Z_g^{a'}Z_g^a1(\hat{\mu}_g^b<\delta)/N^a; \\ T_{2H}^a(\delta) &= \left(\sum_g\rho_g v_g^{b'}Z_g^b(Z_g^{b'}Z_g^b)^{-1}Z_g^{a'}Z_g^a1(\hat{\mu}_g^b\geq\delta)+\sum_g\rho_gZ_g^{a'}v_g^a\right)/N^a; \\ Z_H^a(\delta) &:= Z_{1H}^a(\delta)+Z_{2H}^a(\delta)=\sum_g v_g^{b'}Z_g^b(Z_g^{b'}Z_g^b)^{-1}Z_g^{a'}v_g^a1(\hat{\mu}_g^b\geq\delta)/N^a-\sum_g\rho_gZ_g^{a'}v_g^a1(\hat{\mu}_g^b<\delta)/N^a. \end{aligned}$$

Conforming to notations in Lemma A6 and let $Z^A(\delta)=0$ hence $\hat{A}^a(\delta)=(h^a+T_{1h}^a(\delta)+T_{2h}^a(\delta))^2-2(h^a)^2(H^a)^{-1}(T_{1H}^a(\delta)+T_{2H}^a(\delta))$. Following Lemma A6, to prove the result stated in the above lemma, we just need to show the following seven statements hold with the defined

$S^a(\delta)$.

1) $h^a = O_p(1)$, $H^a = O_p(1)$, 2) $\sup_{\delta\in\Delta} S^a(\delta) = o_p(1)$, 3) $\sup_{\delta\in\Delta}\left(T^a_{1h}(\delta) + T^a_{2h}(\delta)\right) = o_p(1)$,

4) $\sup_{\delta\in\Delta}\left((T^a_{1H}(\delta) + T^a_{2H}(\delta))^2/S^a(\delta)\right) = o_p(1)$, 5) $\sup_{\delta\in\Delta}\left((T^a_{1H}(\delta) + T^a_{2H}(\delta))(T^a_{1h}(\delta) + T^a_{2h}(\delta))/S^a(\delta)\right) = o_p(1)$,

6) $\sup_{\delta\in\Delta}\left(Z^a_H(\delta)/S^a(\delta)\right) = o_p(1)$, 7) $\sup_{\delta\in\Delta}\dfrac{E[\hat{A}^a(\delta)] - \sigma_u^2 H^a - (H^a)^2 S^a(\delta)}{S^a(\delta)} = o_p(1)$.

Step 1: Decomposition of $E[\hat{A}^a(\delta)]$

We have

$$\begin{aligned}E[\hat{A}^a(\delta)] =&\sigma_u^2 H^a + \left(E[(h^a + T^a_{1h}(\delta))^2] - \sigma_u^2 H^a\right) + 2E[(h^a + T^a_{1h}(\delta))T^a_{2h}(\delta)] + E[(T^a_{2h}(\delta))^2]\\ &- 2E[(h^a)^2 T^a_{1H}(\delta)/H^a] - 2E[(h^a)^2 T^a_{2H}(\delta)/H^a]\\ =&\sigma_u^2 H^a + \Delta^a_1(\delta) + \Delta^a_2(\delta) + \Delta^a_3(\delta) + \Delta^a_4(\delta) + \Delta^a_5(\delta).\end{aligned}$$

For the $\Delta^a_1(\delta)$ term,

$$\begin{aligned}\Delta^a_1(\delta) &= E\left[\left(\sum_g \rho_g 1(\hat{\mu}^b_g \geq \delta) Z_g^{a'} u_g^a\right)^2\right]/N^a - \sigma_u^2 H^a\\ &= \sum_g \rho_g^2 E[1(\hat{\mu}^b_g \geq \delta)(Z_g^{a'} u_g^a)^2]/N^a - \sigma_u^2 H^a = \sigma_u^2 \sum_g (\mu^a_g)^2(1 - \Phi^b_{g,\delta})/N^a - \sigma_u^2 H^a\\ &= -\sigma_u^2 \sum_g (\mu^a_g)^2 \Phi^b_{g,\delta}/N^a.\end{aligned}$$

For the $\Delta^a_2(\delta)$ term,

$$\begin{aligned}\Delta^a_2(\delta) =&2\rho_g E\left[\sum_g (Z^a_g)' u^a_g 1(\hat{\mu}^b_g \geq \delta) \sum_g (v^b_g)' Z^b_g ((Z^b_g)' Z^b_g)^{-1} (Z^a_g)' u^a_g 1(\hat{\mu}^b_g \geq \delta)/N\right]\\ =&2\sum_g \rho_g E[((Z^a_g)' u^a_g)^2] E\left[(v^b_g)' Z^b_g ((Z^b_g)' Z^b_g)^{-1} 1(\hat{\mu}^b_g \geq \delta)/N\right]\\ =&2\sigma_u^2 \sigma_v \sum_g \rho_g (Z^a_g)' Z^a_g \phi^b_{g,\delta}/\sqrt{(Z^b_g)' Z^b_g}\\ =&2\sigma_u^2 \sigma_v \sum_g \mu^a_g \phi^b_{g,\delta} \sqrt{\left((Z^a_g)' Z^a_g\right) / \left((Z^b_g)' Z^b_g\right)}.\end{aligned}$$

For the $\Delta_3^a(\delta)$ term,

$$\begin{aligned}
\Delta_3^a(\delta) &= E\left[\left(\sum_g (v_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1} (Z_g^a)' u_g^a 1(\hat{\mu}_g^b \geq \delta)\right)^2\right] / N^a \\
&= \sum_g E\left[\left((v_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1} (Z_g^a)' u_g^a\right)^2 1(\hat{\mu}_g^b \geq \delta)\right] / N^a \\
&= \sum_g E\left[\left((v_g^b)' Z_g^b\right)^2 1(\hat{\mu}_g^b \geq \delta)\right] / ((Z_g^b)' Z_g^b)^2 E\left[(Z_g^a)' u_g^a)^2\right] / N^a \\
&= \sigma_v^2 \sigma_u^2 \sum_g \left(1 - \Phi_{g,\delta}^b + t_{g,\delta}^b \phi_{g,\delta}^b\right) (Z_g^{b'} Z_g^b)^{-1} (Z_g^{a'} Z_g^a) / N^a \\
&= \sigma_v^2 \sigma_u^2 \sum_g \left(1 - \Phi_{g,\delta}^b + t_{g,\delta}^b \phi_{g,\delta}^b\right) / N^a + \sigma_v^2 \sigma_u^2 \sum_g \left(1 - \Phi_{g,\delta}^b + t_{g,\delta}^b \phi_{g,\delta}^b\right) (Z_g^{a'} Z_g^a / (Z_g^{b'} Z_g^b) - 1) / N^a \\
&= \sigma_v^2 \sigma_u^2 \sum_g \left(1 - \Phi_{g,\delta}^b + t_{g,\delta}^b \phi_{g,\delta}^b\right) / N^a + o_p(G/N).
\end{aligned}$$

For the last equality to hold, it suffices to show that $Z_g^{a'} Z_g^a / (Z_g^{b'} Z_g^b) - 1 = O_p(\sqrt{G/N})$ for all groups. Note $Z_g = M_{X_g} \eta_g = \eta_g - P_{X_g} \eta_g$ where η_g is the residual of a linear projection of $\tilde{Z}_g$ onto X_g (i.e. $\tilde{Z}_g = X_g \lambda + \eta_g$). Therefore $Z_g' Z_g = \eta_g' \eta_g - \eta_g' P_{X_g} \eta_g$. Under Assumption 1 we have $E[\eta_{ig}^2] = k_g$ and $E[(\eta_{ig}^2 - k_g)^2] \leq \bar{\Delta}_\eta < \infty$ and $\eta_g' P_{X_g} \eta_g = O_p(E[\eta_g' P_{X_g} \eta_g]) = O_p(1)$ uniformly over all groups. Now by Markov inequality, for any arbitrary $\epsilon > 0$ and pick $C^2 = \bar{\Delta}_\eta \bar{c}$, then

$$\begin{aligned}
P\Big(\sum_i ((\eta_{ig}^a)^2 - k_g) \geq C(N/G)^{1/2}\Big) &\leq \frac{E|\sum_i ((\eta_{ig}^a)^2 - k_g)|^2}{C^2 N/G} \leq \frac{2\sum_i E|(\eta_{ig}^a)^2 - k_g|^2}{C^2 N/G} \\
&= \frac{2 n_g^a E[((\eta_{1g})^2 - k_g)^2]}{C^2 N/G} \leq \frac{1}{C^2} \bar{\Delta}_\eta \bar{c} = \epsilon.
\end{aligned}$$

This implies that $Z_g^{a'} Z_g^a / n_g^a - k_g = O_p(1/\sqrt{N/G})$ for all groups. Similar result holds for the other split of the sample and therefore $Z_g^{a'} Z_g^a / (Z_g^{b'} Z_g^b) - 1 = O_p(\sqrt{G/N})$ for all groups.

For the $\Delta_4^a(\delta)$ term,

$$\begin{aligned}\Delta_4^a(\delta) =& 2E\left[\left(\sum_g \rho_g(Z_g^a)'u_g^a\right)^2 \sum_g \rho_g^2(Z_g^a)'Z_g^a 1(\hat{\mu}_g^b < \delta)\right]/((N^a)^2H^a)\\ =& 2E\left[\left(\sum_g \rho_g(Z_g^a)'u_g^a\right)^2\right]\sum_g(\mu_g^a)^2 E\left[1(\hat{\mu}_g^b < \delta)\right]/((N^a)^2H^a)\\ =& 2\sum_g \rho_g^2 E\left[\left(Z_g^a)'u_g^a\right)^2\right]\sum_g(\mu_g^a)^2\Phi_{g,\delta}^b/((N^a)^2H^a) = 2\sigma_u^2\sum_g(\mu_g^a)^2\Phi_{g,\delta}^b/N^a.\end{aligned}$$

For the $\Delta_5^a(\delta)$ term,

$$\begin{aligned}\Delta_5^a(\delta) =& -2E\left[\left(\sum_g \rho_g(Z_g^a)'u_g^a\right)^2 \sum_g \rho_g(v_g^b)'Z_g^b((Z_g^b)'Z_g^b)^{-1}(Z_g^a)'Z_g^a 1(\hat{\mu}_g^b \geq \delta)\right]/((N^a)^2H^a)\\ & -2E\left[\left(\sum_g \rho_g(Z_g^a)'u_g^a\right)^2 \sum_g \rho_g(Z_g^a)'v_g^a\right]/((N^a)^2H^a)\\ =& -2\sum_g E\left[\left(\rho_g(Z_g^a)'u_g^a\right)^2\right]\sum_g E\left[\rho_g(v_g^b)'Z_g^b((Z_g^b)'Z_g^b)^{-1}(Z_g^a)'Z_g^a 1(\hat{\mu}_g^b \geq \delta)\right]/((N^a)^2H^a)\\ =& -2\sigma_u^2\sigma_v\sum_g \mu_g^a\phi_{g,\delta}^b\sqrt{((Z_g^a)'Z_g^a)/((Z_g^b)'Z_g^b)}/N^a.\end{aligned}$$

Step 2: Determine $S^a(\delta)$

Collecting the leading terms from $\Delta_1^a(\delta)$ to $\Delta_5^a(\delta)$ we get

$$\sigma_v^2\sigma_u^2\sum_g\left(1-\Phi_{g,\delta}^b + t_{g,\delta}^b\phi_{g,\delta}^b\right)/N^a + \sigma_u^2\sum_g(\mu_g^a)^2\Phi_{g,\delta}^b/N^a,$$

which is the $(H^a)^2S^a(\delta)$ term defined in Lemma A7. Modifying the convergence results derived in Lemma A4 for the subsampled analysis, it is easy to show that $\sup_{\delta\in\Delta} S^a(\delta) = O_p(G/N)$. In addition, $\sum_{g\in\mathcal{G}_{+,s}}\left(\Phi_{g,\delta} - t_{g,\delta}\phi_{g,\delta}\right)/N = O_p(G^3/N^3)$. Therefore, $\inf_{\delta\in\Delta} S^a(\delta) \geq \sigma_v^2\sigma_u^2 bG/N + o_p(G/N)$ and any terms of order $o_p(G/N)$ is dominated by $S^a(\delta)$ uniformly over $\delta \in \Delta$.

Step 3: Prove statements corresponding to (1) - (6) in Lemma A6

For statement (1), $h^a = O_p(1)$ and $H^a = O_p(1)$ are shown in the proof of Lemma 1.

For statement (2), since we have $\sup_{\delta\in\Delta} S^a(\delta) = O_p(G/N)$, hence $\sup_{\delta\in\Delta} S^a(\delta) = o_p(1)$ as $G/N \to 0$.

For statement (3), note that $\sup_{\delta\in\Delta} T^a_{1h}(\delta) = o_p(1)$ and $\sup_{\delta\in\Delta} T^a_{2h}(\delta) = o_p(1)$ by the Markov inequality and the facts that $E[|T^a_{1h}(\delta)|] \leq \sqrt{E[(T^a_{1h}(\delta))^2]} = O_p(\sqrt{G/N})$ for all $\delta \in \Delta$, and $E[|T^a_{2h}(\delta)|] \leq \sqrt{E[(T^a_{2h}(\delta))^2]} = O_p(\sqrt{G/N})$ for all $\delta \in \Delta$.

To prove (4), notice that $\sup_{\delta\in\Delta} T^a_{1H}(\delta) = O_p(G/N)$ by the Markov inequality. Furthermore we have $E[|T^a_{1H}(\delta)|] = E\Big[\sum_g \rho_g^2 Z_g^{a\prime} Z_g^a 1(\hat{\mu}_g^b < \delta)\Big]/N^a = \sum_g (\mu_g^a)^2 \Phi^b_{g,\delta}/N^a$. Applying results in Lemma A4 proves statement (4).

For $T^a_{2H}(\delta)$, notice that its second component is free from δ and is $O_p(1/\sqrt{N})$ by the central limit theorem, and its first component is $O_p(G/N)$ uniformly over $\delta \in \Delta$ by the Markov inequality and the fact that

$$
\begin{aligned}
&E\Big[\Big(\sum_g \rho_g (v_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1} (Z_g^a)' Z_g^a 1(\hat{\mu}_g^b > \delta)\Big)^2\Big]/(N^a)^2 \\
=&\sigma_v^2 \left(\sum_g (\mu_g^b)^2 (1 - \Phi^b_{g,\delta} + t^b_{g,\delta}\phi^b_{g,\delta}) + \left(\sum_g \mu_g^b \phi^b_{g,\delta}\right)^2 - \sum_g (\mu_g^b)^2 (\phi^b_{g,\delta})^2\right) ((Z_g^a)' Z_g^a/(Z_g^b)' Z_g^b)^2/(N^a)^2 \\
=&O_p(G^2/N^2).
\end{aligned}
$$

Since each of $1/N$, G^2/N^2, and $G/N/\sqrt{N}$ is of order $o_p(G/N)$, hence $\sup_{\delta\in\Delta}\Big((T^a_{1H} + T^a_{2H})^2/S^a(\delta)\Big) = o_p(1)$.

For statement (5), note that by statements (3) and (4) $\sup_{\delta\in\Delta}\Big(T^a_{1h}(\delta) + T^a_{2h}(\delta)\Big) = o_p(1)$, and $\sup_{\delta\in\Delta}\Big(T^a_{1H}(\delta) + T^a_{2H}(\delta)\Big) = O_p(G/N)$. Therefore $\sup_{\delta\in\Delta}\Big((T^a_{1h} + T^a_{2h})(T^a_{1H} + T^a_{2H})/S^a(\delta)\Big) = o_p(1)$.

Lastly, for statement (6), notice that the first term of $Z^a_H(\delta)$ is of order $O_p(\frac{\sqrt{G}}{N})$ by the

Markov inequality, the Cauchy-Schwarz inequality and the facts that for all $\delta \in \Delta$,

$$\begin{aligned}
E[(Z_{1H}^a(\delta))^2] &= E\left[\left(\sum_g (v_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1} (Z_g^a)' v_g^a 1(\hat{\mu}_g^b > \delta)\right)^2\right] /(N^a)^2 \\
&= \sum_g E\left[((v_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1})^2 1(\hat{\mu}_g^b > \delta)\right] E\left[(Z_g^a)' v_g^a)^2\right] /(N^a)^2 \\
&= \sigma_v^4 \sum_g (1 - \Phi_{g,\delta}^b + t_{g,\delta}^b \phi_{g,\delta}^b)(Z_g^a)' Z_g^a / ((Z_g^b)' Z_g^b) / (N^a)^2 = O_p(G/N^2), \\
E[(Z_{2H}^a(\delta))^2] &= E\left[\left(\sum_g \rho_g (Z_g^a)' v_g^a 1(\hat{\mu}_g^b < \delta)\right)^2\right] /(N^a)^2 = \sigma_v^2 \sum_{g \in G} (\mu_g^a)^2 \Phi_{g,\delta}^b / (N^a)^2 = O_p(G^2/N).
\end{aligned}$$

Putting together, we have $\sup_{\delta \in \Delta} \left(Z_H^a(\delta) / S^a(\delta) \right) = o_p(1)$.

Following the three steps, the lemma is hence proven. □

Finally, we are going to collect the asymptotic limits of both split-sample estimators and derive the asymptotic limit of the repeated split-sample select-and-interact estimator.

Note that by Lemma A7, we know that

$$\begin{aligned}
\sqrt{N^a}(\hat{\beta}_{sssel,int}^a(\delta) - \beta) &= (\hat{H}_\delta^a)^{-1} \hat{h}_\delta^a \\
&= (H^a)^{-1} \left(h^a + T_{1h}^a(\delta) + T_{2h}^a(\delta) - (T_{1H}^a(\delta) + T_{2H}^a(\delta))(H^a)^{-1} h^a \right) + o_p(G/N),
\end{aligned}$$

and a similar result holds for $\hat{\beta}_{sssel,int}^b(\delta)$.

Since $n_g - 1 \leq 2n_g^a \leq n_g + 1$, we know that $|1/N_a - 2/N| = O(G/N^2)$ and a similar result holds for N^b. Furthermore, since $\hat{H}_\delta^a$, $\hat{h}_\delta^a$, $\hat{H}_\delta^b$, $\hat{h}_\delta^b$ are all $O_p(1)$, we know that

$$\begin{aligned}
\sqrt{N}(\hat{\beta}_{ssel,int}(\delta) - \beta) =& \frac{1}{2} \left(\sqrt{N/N^a} \sqrt{N^a} (\hat{\beta}_{sssel,int}^a(\delta) - \beta) + \sqrt{N/N^b} \sqrt{N^b} (\hat{\beta}_{sssel,int}^b(\delta) - \beta) \right) \\
=& \left(h^a + T_{1h}^a(\delta) + T_{2h}^a(\delta) - (T_{1H}^a(\delta) + T_{2H}^a(\delta))(H^a)^{-1} h^a \right) / H^a / \sqrt{2} \\
&+ \left(h^b + T_{1h}^b(\delta) + T_{2h}^b(\delta) - (T_{1H}^b(\delta) + T_{2H}^b(\delta))(H^b)^{-1} h^b \right) / H^b / \sqrt{2} \\
&+ o_p(G/N).
\end{aligned}$$

Denote $\hat{Q}^a(\delta) = E[(h^a + T_{1h}^a(\delta) + T_{2h}^a(\delta) - (T_{1H}^a(\delta) + T_{2H}^a(\delta))(H^a)^{-1} h^a) / H^a$ and denote

$\hat{Q}^b(\delta)$ similarly. Notice that

$$
\begin{aligned}
&\left(h^a + T^a_{1h}(\delta) + T^a_{2h}(\delta) - (T^a_{1H}(\delta) + T^a_{2H}(\delta))(H^a)^{-1}h^a\right)\left(h^b + T^b_{1h}(\delta) + T^b_{2h}(\delta) - (T^b_{1H}(\delta) + T^b_{2H}(\delta))(H^b)^{-1}h^b\right)\\
=&(h^a + T^a_{1h}(\delta) + T^a_{2h}(\delta))\left(h^b + T^b_{1h}(\delta) + T^b_{2h}(\delta)\right) - (h^a + T^a_{1h}(\delta) + T^a_{2h}(\delta))\,(T^b_{1H}(\delta) + T^b_{2H}(\delta))(H^b)^{-1}h^b\\
&- (T^a_{1H}(\delta) + T^a_{2H}(\delta))(H^a)^{-1}h^a\left(h^b + T^b_{1h}(\delta) + T^b_{2h}(\delta)\right)\\
&+ (T^a_{1H}(\delta) + T^a_{2H}(\delta))(H^a)^{-1}h^a(T^b_{1H}(\delta) + T^b_{2H}(\delta))(H^b)^{-1}h^b\\
=&(h^a + T^a_{1h}(\delta) + T^a_{2h}(\delta))\left(h^b + T^b_{1h}(\delta) + T^b_{2h}(\delta)\right)\\
&- h^a(T^b_{1H}(\delta) + T^b_{2H}(\delta))(H^b)^{-1}h^b - (T^a_{1H}(\delta) + T^a_{2H}(\delta))(H^a)^{-1}h^ah^b\\
&+ o_p(G/N),
\end{aligned}
$$

where the last equality holds since in the proof of Lemma A7 we also showed that $h^a = O_p(1)$, $H^a = O_p(1)$ and $\sup_{\delta\in\Delta} T^a_{1h}(\delta) + T^a_{2h}(\delta) = O_p(\sqrt{G/N})$ and $\sup_{\delta\in\Delta} T^a_{1H}(\delta) + T^a_{2H}(\delta) = o_p(\sqrt{G/N})$ and similar results hold for subsample b. Denote $\hat{Q}^{ab}(\delta)$ such that

$$
\begin{aligned}
\hat{Q}^{ab}(\delta)H^aH^b =&(h^a + T^a_{1h}(\delta) + T^a_{2h}(\delta))\left(h^b + T^b_{1h}(\delta) + T^b_{2h}(\delta)\right)\\
&- h^a(T^b_{1H}(\delta) + T^b_{2H}(\delta))(H^b)^{-1}h^b - (T^a_{1H}(\delta) + T^a_{2H}(\delta))(H^a)^{-1}h^ah^b.
\end{aligned}
$$

Then we know

$$
N(\hat{\beta}_{sssel,int}(\delta) - \beta)^2 = \hat{Q}^a(\delta)/2 + \hat{Q}^a(\delta)/2 + \hat{Q}^{ab}(\delta) + o_p(G/N).
$$

where the form of $E[\hat{Q}^a(\delta)]$ is derived in Lemma A7, and a similar result holds for $E[\hat{Q}^b(\delta)]$. For the last term, recognizing that $E[h^a(T^b_{1H}(\delta) + T^b_{2H}(\delta))h^b] = 0$, we know that

$$
\begin{aligned}
E[\hat{Q}^{ab}(\delta)]H^aH^b &= E[(h^a + T^a_{1h}(\delta) + T^a_{2h}(\delta))\left(h^b + T^b_{1h}(\delta) + T^b_{2h}(\delta)\right)]\\
&= E[(T^a_{1h}(\delta) + T^a_{2h}(\delta))(T^b_{1h}(\delta) + T^b_{2h}(\delta))]
\end{aligned}
$$

as $E[h^ah^b] = E[h^aT^b_{1h}(\delta)] = E[h^aT^b_{2h}(\delta)] = E[T^a_{1h}(\delta)h^b] = E[T^a_{2h}(\delta)h^b] = 0$. For the last

expectation, we know that

$$
\begin{aligned}
&E[(T^a_{1h}(\delta)+T^a_{2h}(\delta))(T^b_{1h}(\delta)+T^b_{2h}(\delta))]\\
=&\sum_g E[(v^a_g)'P_{Z^a_g}u^a_g 1(\hat\mu^a_g\geq\delta)]E[(v^b_g)'P_{Z^b_g}u^b_g 1(\hat\mu^b_g\geq\delta)]/\sqrt{N^aN^b}\\
&+\sum_g \rho_g^2 E[1(\hat\mu^a_g<\delta)(Z^a_g)'u^a_g]E[1(\hat\mu^b_g<\delta)(Z^b_g)'u^b_g]/\sqrt{N^aN^b}\\
&-\sum_g \rho_g E[(v^a_g)'P_{Z^a_g}u^a_g 1(\hat\mu^a_g\geq\delta)]E[1(\hat\mu^b_g<\delta)(Z^b_g)'u^b_g]/\sqrt{N^aN^b}\\
&-\sum_g \rho_g E[(v^b_g)'P_{Z^b_g}u^b_g 1(\hat\mu^b_g\geq\delta)]E[1(\hat\mu^a_g<\delta)(Z^a_g)'u^a_g]/\sqrt{N^aN^b}\\
=&\sum_g \sigma^2_{uv}\left(1-\Phi^a_{g,\delta}+t^a_{g,\delta}\phi^a_{g,\delta}\right)\left(1-\Phi^b_{g,\delta}+t^b_{g,\delta}\phi^b_{g,\delta}\right)/(N/2)+\frac{\sigma^2_{uv}}{\sigma^2_v}\sum_g \mu^a_g\mu^b_g\phi^a_{g,\delta}\phi^b_{g,\delta}/(N/2)\\
&+\frac{\sigma^2_{uv}}{\sigma_v}\sum_g\left[\mu^b_g\left(1-\Phi^a_{g,\delta}+t^a_{g,\delta}\phi^a_{g,\delta}\right)\phi^b_{g,\delta}+\mu^a_g\left(1-\Phi^b_{g,\delta}+t^b_{g,\delta}\phi^b_{g,\delta}\right)\phi^a_{g,\delta}\right]/(N/2)+o_p(G/N)\\
=&2\sigma^2_{uv}\sum_g\left(1-\Phi^a_{g,\delta}+\frac{\delta}{\sigma_v}\phi^a_{g,\delta}\right)\left(1-\Phi^b_{g,\delta}+\frac{\delta}{\sigma_v}\phi^b_{g,\delta}\right)/N+o_p(G/N).
\end{aligned}
$$

Further, since $\left|\Phi(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v})-\Phi(\frac{\delta-\mu^a_g}{\sigma_v})\right|\leq\left|\mu_g/\sqrt{2}-\mu^a_g\right|\phi\left(\frac{\delta-\mu^*_g}{\sigma_v}\right)$ for a μ^*_g between $\mu_g/\sqrt{2}$ and μ^a_g, by Cauchy-Schwarz inequality we have the following results:

$$
\begin{aligned}
E\left|\Phi(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v})-\Phi(\frac{\delta-\mu^a_g}{\sigma_v})\right|&\leq\sqrt{E\left[\left(\mu_g/\sqrt{2}-\mu^a_g\right)^2\right]}\sqrt{E[\phi^2(\frac{\delta-\mu^*_g}{\sigma_v})]}\\
&\leq\rho_g^2 n_g\cdot\sqrt{E\left[\left(\sqrt{Z_g'Z_g/n_g}-\sqrt{(Z^a_g)'Z^a_g/n^a_g}\right)^2\right]}\sqrt{E[\phi^2(\frac{\delta-\mu^*_g}{\sigma_v})]}\\
&\lesssim\sqrt{E\left[\left(\sqrt{Z_g'Z_g/n_g}-\sqrt{(Z^a_g)'Z^a_g/n^a_g}\right)^2\right]}\lesssim(G/N)^{-1/4}
\end{aligned}
$$

and

$$
\left|\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\phi(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v})-\frac{\delta-\mu^a_g}{\sigma_v}\phi(\frac{\delta-\mu^a_g}{\sigma_v})\right|\leq\left|\mu_g/\sqrt{2}-\mu^a_g\right|\left|\frac{\delta-\mu^*_g}{\sigma_v}\right|\phi\left(\frac{\delta-\mu^*_g}{\sigma_v}\right)\lesssim(G/N)^{-1/4}
$$

and

$$
\begin{aligned}
\left|\mu_g/\sqrt{2}\phi(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v})-\mu^a_g\phi(\frac{\delta-\mu^a_g}{\sigma_v})\right|&\leq\left|\mu_g/\sqrt{2}-\mu^a_g\right|\phi\left(\frac{\delta-\mu^*_g}{\sigma_v}\right)+\left|\mu_g/\sqrt{2}-\mu^a_g\right|\left|\frac{\delta-\mu^*_g}{\sigma_v}\right|\phi\left(\frac{\delta-\mu^*_g}{\sigma_v}\right)\\
&\lesssim(G/N)^{-1/4}
\end{aligned}
$$

Therefore, we know that

$$E[\hat{Q}^{ab}(\delta)] = 2\sigma_{uv}^2 \sum_g \left(1 - \Phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right) + \frac{\delta}{\sigma_v}\phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right)\right)^2 / H^a / H^b / N + o_p(G/N).$$

Putting together, we get that

$$\begin{aligned}
&E[\hat{Q}^a(\delta)|\tilde{Z}, X] + E[\hat{Q}^b(\delta)|\tilde{Z}, X] \\
=&\sigma_u^2(1/H^a + 1/H^b)/2 \\
&+ \sigma_u^2\sigma_v^2 \sum_g \left(1 - \Phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right) + \left(\frac{\delta - \mu_g/\sqrt{n_g}}{\sigma_v}\right)\phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right)\right)\left(1/(H^a)^2 + 1/(H^b)^2\right)/N \\
&+ \sigma_u^2 \sum_g \mu_g^2 \Phi\left(\frac{\delta - \mu_g/\sqrt{2}}{\sigma_v}\right)\left(1/(H^a)^2 + 1/(H^b)^2\right)/N.
\end{aligned}$$

Moreover under Assumption 1 we obtain $\frac{1}{N^a}\sum_g (Z_g^a)'Z_g^a = \frac{1}{N^a}\sum_g n_g^a((Z_g^a)'Z_g/n_g^a - k_g) + \frac{1}{N^a}\sum_g n_g^a k_g = \bar{k} + O_p(\frac{G}{N}\sqrt{\frac{G}{N}}) = \bar{k} + o_p(\frac{G}{N})$. A similar argument yields $\sum_g Z_g'Z_g/N = \bar{k} + o_p(G/N)$. Put together, we have $|H^a - H| \leq o_p(G/N)$. A similar result holds for subsample b. Collecting all terms that contribute to the $N(\hat{\beta}_{sssel,int}(\delta) - \beta)^2$ leads to the formulation of $S_{sssel,int}(\delta)$ stated in the theorem.

□

Proof of Corollary 1

Proof. Recall the definition of $S_{sssel,int}(\delta)$, $A_{sssel,int}(\delta)$, $B_{sssel,int}(\delta)$, and $C_{sssel,int}(\delta)$ in Theorem 2. In this proof, we omit the subscript for notational simplicity.

Recall the definition of $A_{sssel,int}(\delta)$, $B_{sssel,int}(\delta)$, and $C_{sssel,int}(\delta)$ in Theorem 2. In this proof, we omit the subscript for notational simplicity.

First, we notice that $A(\delta) \geq 0$ as $1 - \Phi(x) + x\phi(x) \geq 0$ and $B(\delta) \geq 0$ as $\Phi(x)$ is nonnegative. Decompose $C(\delta)$ to $C_1(\delta) + C_2(\delta)$ where $C_1(\delta) = 2\frac{\sigma_{uv}^2}{N}\sum_g \left(1 - \Phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right) + \frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right)\right)^2$ and $C_2(\delta) = 4\frac{\sigma_{uv}^2}{N}\left(1-\Phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right)+\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right)\right)\frac{\mu_g/\sqrt{2}}{\sigma_v}\phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right) + \frac{\sigma_{uv}^2}{\sigma_v^2 N}\mu_g^2\phi^2\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right)$. It is clear that both $C_1(\delta)$ and $C_2(\delta)$ are non-negative as well.

In addition, we have $\nabla_\delta A(\delta) = -\frac{2\sigma_u^2\sigma_v}{N}\sum_g \left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right)^2 \phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right) \leq 0$, $\nabla_\delta B(\delta) = \frac{\sigma_u^2}{\sigma_v N}\sum_g \mu_g^2\phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right) \geq 0$

and $\nabla_\delta C_1(\delta) = -2\frac{\sigma_{uv}^2}{\sigma_v N}\sum_g \left(1-\Phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right)+\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right)\right)\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right)^2\phi\left(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right) \leq 0$. The sign of $\nabla_\delta C_2(\delta)$ is generally ambiguous but it is easy to show that $\nabla_\delta C_2(\delta) \geq 0$ when $\delta - \mu_g/\sqrt{2} \leq 0$.

Denote $L(\delta) \equiv A(\delta) + B(\delta) + C_1(\delta) + C_2(\delta)$. Next, we use superscripts to indicate that a term is evaluated for particular set of groups. For example, contributions from the groups in the set $\mathcal{G}_0$ are denoted as $L^0(\delta)$ and contributions from the groups in the set $\mathcal{G}_{+,w}$ are denoted as $L^{+,w}(\delta)$. We have $L^0(\delta) \equiv A^0(\delta) + B^0(\delta) + C_1^0(\delta) + C_2^0(\delta)$ with $A^0(\delta) = 2\frac{\sigma_u^2\sigma_v^2}{N}\sum_{g\in\mathcal{G}_0}\left(1-\Phi\left(\frac{\delta}{\sigma_v}\right)+\frac{\delta}{\sigma_v}\phi\left(\frac{\delta}{\sigma_v}\right)\right)^2$, $C_1^0(\delta) = 2\frac{\sigma_{uv}^2\sigma_v^2}{N}\sum_{g\in\mathcal{G}_0}\left(1-\Phi\left(\frac{\delta}{\sigma_v}\right)+\frac{\delta}{\sigma_v}\phi\left(\frac{\delta}{\sigma_v}\right)\right)^2$, and $B^0(\delta) = C_2^0(\delta) = 0$. To prove the Corollary, we show in sequence the following results:

1) As $N \to \infty$, $L^0(\delta_N) = o(\frac{G}{N})$ for any sequence $\delta_N \to +\infty$.

2) For any $\delta \geq 0$, $L^{+,w}(\delta) = o_p(\frac{G}{N})$ when $G_2/G \to 0$.

3) $L(\delta) \geq 2b\sigma_u^2\sigma_v^2(1+\rho_{uv}^2)\frac{G}{N} + o_p(\frac{G}{N})$ for $\forall\delta \geq 0$.

Proof of 1): For any sequence $\delta_N \to \infty$, we have

$$
\begin{aligned}
L^0(\delta_N) =& A^0(\delta_N) + C_1^0(\delta_N) \\
=& 2\sigma_u^2\sigma_v^2\frac{G_0}{N}\cdot\left(1-\Phi\left(\frac{\delta_N}{\sigma_v}\right)+\frac{\delta_N}{\sigma_v}\phi\left(\frac{\delta_N}{\sigma_v}\right)\right) + 2\sigma_{uv}^2\frac{G_0}{N}\cdot\left(1-\Phi\left(\frac{\delta_N}{\sigma_v}\right)+\frac{\delta_N}{\sigma_v}\phi\left(\frac{\delta_N}{\sigma_v}\right)\right)^2 \\
\leq& 2\sigma_u^2\sigma_v^2\frac{G_0}{N}\cdot\left(1/\left(\frac{\delta_N}{\sigma_v}\right)+\frac{\delta_N}{\sigma_v}\right)\phi\left(\frac{\delta_N}{\sigma_v}\right) + 2\sigma_{uv}^2\frac{G_0}{N}\cdot\left(\left(1/\left(\frac{\delta_N}{\sigma_v}\right)+\frac{\delta_N}{\sigma_v}\right)\phi\left(\frac{\delta_N}{\sigma_v}\right)\right)^2 \\
\leq& \sigma_u^2\sigma_v^2\frac{G_0}{N}\cdot\left(1/\left(\frac{\delta_N}{\sigma_v}\right)^3+1/\left(\frac{\delta_N}{\sigma_v}\right)\right) + \sigma_{uv}^2\frac{G_0}{N}\cdot\left(\left(1/\left(\frac{\delta_N}{\sigma_v}\right)^3+1/\left(\frac{\delta_N}{\sigma_v}\right)\right)\right)^2 \\
=& o(G/N).
\end{aligned}
$$

where the first inequality holds as $1-\Phi(x) \leq \phi(x)/x$ for any $x > 0$ and the second holds as $\phi(x) \leq 1/x^2/\sqrt{2\pi}$.

Proof of 2): For the weak groups, for any $\delta \geq 0$, we have that

$$
B^{+,w}(\delta) \leq \frac{\sigma_u^2}{N}\sum_{g\in\mathcal{G}_{+,w}}\mu_g^2 = \frac{G_{+,w}}{G}O_p(G/N)
$$

and

$$\begin{aligned}A^{+,w}(\delta)+C_1^{+,w}(\delta)\leq&\frac{\sigma_u^2\sigma_v^2}{N}\sum_{g\in\mathcal{G}_{+,w}}\left(1/\left|\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right|^3+1/\left|\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right|\right)\\&+\frac{\sigma_{uv}^2}{N}\sum_{g\in\mathcal{G}_{+,w}}\left(\left(1/\left|\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right|^3+1/\left|\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\right|\right)\right)^2\\=&\frac{G_{+,w}}{G}O_p(G/N),\end{aligned}$$

Similarly,

$$\begin{aligned}C_2^{+,w}(\delta)\leq&\sum_{g\in\mathcal{G}_{+,w}}2\frac{\sigma_{uv}^2}{N}\frac{\mu_g/\sqrt{2}}{\sigma_v}\phi\Big(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\Big)+\sum_{g\in\mathcal{G}_{+,w}}\frac{\sigma_{uv}^2}{N\sigma_v^2}\mu_g^2\phi^2\Big(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\Big)\\=&\frac{G_{+,w}}{G}O_p(G/N).\end{aligned}$$

When $G_{+,w}/G\to 0$, we have $L^{+,w}(\delta)=o_p(G/N)$ for any $\delta\geq 0$. The second statement is hence proven.

Proof of 3):

Let $\mathcal{U}\equiv 2\sigma_u^2\sigma_v^2+2\sigma_{uv}^2=2\sigma_u^2\sigma_v^2(1+\rho_{uv}^2)$. Focus on the contribution of one group $g\in\mathcal{G}_{+,s}$ to the terms $B^{+,s}(\delta)$, $A^{+,s}(\delta)$ and $C_1^{+,s}(\delta)$. For that group g, let δ_g^* satisfy that $\sigma_u^2\mu_g^2\Phi(\frac{\delta_g^*-\mu_g/\sqrt{2}}{\sigma_v})=\mathcal{U}$. Then it is easy to derive that $\Phi\Big(\frac{\delta_g^*-\mu_g/\sqrt{2}}{\sigma_v}\Big)=\frac{\mathcal{U}}{\sigma_u^2\mu_g^2}$, or that $\delta_g^*=\mu_g/\sqrt{2}+\sigma_v\Phi^{-1}\Big(\frac{\mathcal{U}}{\sigma_u^2\mu_g^2}\Big)$.

For any $\delta>\delta_g^*$, since each term in $B^{+,s}(\delta)$ increases in δ, we know that $\frac{\sigma_u^2}{N}\mu_g^2\Phi\Big(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\Big)\geq\frac{\mathcal{U}}{N}\geq\frac{G_{+,s}}{G}\frac{\mathcal{U}}{N}$. For any $0\leq\delta\leq\delta^*$, notice that

$$\begin{aligned}&\left|2\frac{\sigma_u^2\sigma_v^2}{N}\mathcal{A}_\delta+2\frac{\sigma_{uv}^2}{N}\mathcal{A}_\delta^2-\frac{G_{+,s}}{G}\frac{\mathcal{U}}{N}\right|\leq 2\frac{\sigma_u^2\sigma_v^2}{N}\left|\mathcal{A}_\delta-1\right|+2\frac{\sigma_{uv}^2}{N}\left|\mathcal{A}_\delta^2-1\right|\\\leq&2\frac{\sigma_u^2\sigma_v^2}{N}\left(1-\mathcal{A}_\delta\right)+4\frac{\sigma_{uv}^2}{N}\left(1-\mathcal{A}_\delta\right)\leq 6\frac{\sigma_u^2\sigma_v^2}{N}\left(1-\mathcal{A}_\delta\right)\\\leq&6\frac{\sigma_u^2\sigma_v^2}{N}\left(1-\mathcal{A}_\delta\right)\leq 6\frac{\sigma_u^2\sigma_v^2}{N}(1-\mathcal{A}_{\delta_g^*})\\=&6\frac{\sigma_u^2\sigma_v^2}{N}\left(\Phi(x_g^*)-x_g^*\phi\Big(x_g^*\Big)\right)\\=&O_p\left(\frac{1}{N}E\left[\Phi(x_g^*)\right]\right)+O_p\left(\frac{1}{N}E\left[|x_g^*|\phi\Big(x_g^*\Big)\right]\right)=o_p(G/N),\end{aligned}$$

where $\mathcal{A}_\delta = 1 - \Phi\Big(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\Big) + \frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\phi\Big(\frac{\delta-\mu_g/\sqrt{2}}{\sigma_v}\Big)$ which decreases with δ and let $x_g^* = (\delta_g^* - \mu_g/\sqrt{2})/\sigma_v = \Phi^{-1}\Big(\frac{\mathcal{U}}{\sigma_u^2\mu_g^2}\Big)$ and $A_{\delta_g^*} = 1 - \Phi(x_g^*) + x_g^*\phi(x_g^*) \leq 1$.

The last equality follows as

$$\begin{aligned}
E\left[\Phi(x_g^*)\right] =& E\left[\Phi(x_g^*)1\left(\mu_g > \rho_g\sqrt{k_g n_g/2}\right)\right] + P(\mu_g \leq \rho_g\sqrt{k_g n_g/2}) \\
=& E\left[\frac{\mathcal{U}}{\sigma_u^2\mu_g^2}1\left(\mu_g > \rho_g\sqrt{k_g n_g/2}\right)\right] + P(\mu_g \leq \rho_g\sqrt{k_g n_g/2}) \\
\leq& \frac{2\sigma_v^2(1+\rho_{uv}^2)}{\rho_g^2 k_g n_g/2} + P\left(Z_g' Z_g/n_g \leq k_g/2\right) \lesssim G/N,
\end{aligned}$$

and

$$\begin{aligned}
E\left[|x_g^*|\phi\Big(x_g^*\Big)\right] \leq& E\left[|x_g^*|\phi\Big(x_g^*\Big)1\left(\mu_g > \rho_g\sqrt{k_g n_g/2}\right)\right] + P\left(\mu_g \leq \rho_g\sqrt{k_g n_g/2}\right) \\
\leq& E\left[|x_g^{**}|\phi\left(x_g^{**}\right)\right] + P\left(Z_g' Z_g/n_g \leq k_g/2\right) \\
\leq& E\left[1/|x_g^{**}|/\sqrt{2\pi}\right] + P\left(Z_g' Z_g/n_g \leq k_g/2\right) \\
\leq& 1/\left|\Phi^{-1}\left(\frac{2\sigma_v^2(1+\rho_{uv}^2)}{\bar{\rho}^2\bar{k}\bar{c}N/G/2}\right)\right| + P\left(Z_g' Z_g/n_g \leq k_g/2\right) = o(1)
\end{aligned}$$

where $x_g^{**} = \Phi^{-1}\left(\frac{2\sigma_v^2(1+\rho_{uv}^2)}{\rho_g^2 k_g n_g/2}\right)$ and last convergence result holds uniformly over g.

Therefore, we know that

$$2\frac{\sigma_u^2\sigma_v^2}{N}\mathcal{A}_\delta + 2\frac{\sigma_{uv}^2}{N}\mathcal{A}_\delta^2 = \frac{G_{+,s}}{G}\frac{\mathcal{U}}{N} + o_p(G/N).$$

Therefore we know that for any $\delta \geq 0$, the contribution from the g group to $B^{+,s}(\delta) + A^{+,s}(\delta) + C_1^{+,s}(\delta)$ is lower bounded by

$$\frac{G_{+,s}}{G}\frac{\mathcal{U}}{N}.$$

Now we can make this argument for all groups $g \in \mathcal{G}_{+,s}$, therefore, since $G_{+,s}/G \to b$ under Assumption 1, we know that $L(\delta) \geq 2b\sigma_u^2\sigma_v^2(1+\rho_{uv}^2)\frac{G}{N} + o_p(\frac{G}{N})$ for $\forall\delta \geq 0$.

□

Proof of Theorem 3

Proof. To establish the result, we first prove the following two statements:

i) $\min_{g\in\mathcal{G}_{+,s}} \hat{\mu}_g^2/\kappa_{G,N}$ diverges with probability approaching 1,

ii) $\max_{g\in\mathcal{G}_{+,s}^c} \hat{\mu}_g^2/\kappa_{G,N}$ goes to zero with probability approaching 1.

To prove i), note

$$\min_{g\in\mathcal{G}_{+,s}} \hat{\mu}_g^2 = \inf_{g\in\mathcal{G}_{+,s}} \{\mu_g^2 + \hat{\mu}_g^2 - \mu_g^2\} \geq \min_{g\in\mathcal{G}_{+,s}} \mu_g^2 - \max_{g\in\mathcal{G}_{+,s}} |\hat{\mu}_g^2 - \mu_g^2|.$$

For the second term on the right hand side, we have

$$\begin{aligned}
&P(\max_{g\in\mathcal{G}_{+,s}} |\hat{\mu}_g^2 - \mu_g^2| > \epsilon) \leq \sum_{g\in\mathcal{G}_{+,s}} P\left(|\hat{\mu}_g^2 - \mu_g^2| > \epsilon\right) \leq G_{+,s} \max_{g\in\mathcal{G}_{+,s}} P\left(|\hat{\mu}_g^2 - \mu_g^2| > \epsilon\right) \\
&= G_{+,s} \max_{g\in\mathcal{G}_{+,s}} P\left(\left|2\rho_g Z_g' v_g + \left(\frac{Z_g' v_g}{\sqrt{Z_g' Z_g}}\right)^2\right| > \epsilon\right) \\
&\leq G \max_{g\in\mathcal{G}_{+,s}} P\left(|2\rho_g Z_g' v_g| + \frac{(Z_g' v_g)^2}{Z_g' Z_g} > \epsilon\right) \\
&= G \max_{g\in\mathcal{G}_{+,s}} \Big\{P\Big(|2\rho_g Z_g' v_g| + \frac{(Z_g' v_g)^2}{Z_g' Z_g} > \epsilon, |2\rho_g Z_g' v_g| > \frac{(Z_g' v_g)^2}{Z_g' Z_g}\Big) \\
&+ P\Big(|2\rho_g Z_g' v_g| + \frac{(Z_g' v_g)^2}{Z_g' Z_g} > \epsilon, |2\rho_g Z_g' v_g| \leq \frac{(Z_g' v_g)^2}{Z_g' Z_g}\Big)\Big\} \\
&\leq G \max_{g\in\mathcal{G}_{+,s}} \Big\{P\Big(|2\rho_g Z_g' v_g| > \epsilon/2\Big) + P\Big(|2\rho_g Z_g' v_g| \leq \frac{(Z_g' v_g)^2}{Z_g' Z_g}\Big)\Big\}.
\end{aligned}$$

Since $v_{ig} \sim \mathcal{N}(0, \sigma_v^2)$, apply the tail bound for Gaussian random variable,[5]

$$P\left(\left|2\rho_g Z_g' v_g\right| > \epsilon/2\right) = P\left(\left|\frac{1}{\sigma_v\sqrt{k_g}\sqrt{n_g}} Z_g' v_g\right| > \frac{\epsilon/2}{2\rho_g\sqrt{n_g}\sigma_v\sqrt{k_g}}\right) \leq 2\exp\left(-\frac{1}{2}\frac{\epsilon^2/4}{4\rho_g^2 n_g \sigma_v^2 k_g}\right),$$

$$P\Big(|2\rho_g Z_g' v_g| \leq \frac{(Z_g' v_g)^2}{Z_g' Z_g}\Big) = P\Big(|(Z_g' Z_g)^{-1} Z_g' v_g/2\rho_g| \geq 1\Big) \leq 2\exp(-\frac{1}{2}\rho_g^2 k_g n_g/\sigma_v^2).$$

Therefore, under Assumption 1 and the rate condition $(G\log G)/N \to 0$, we have that when $\epsilon = C_1\sqrt{\frac{N}{G}\log G}$ with $C_1^2 > 32\bar{c}\bar{\rho}^2\sigma_v^2\bar{k}$,

$$\begin{aligned}
P(\max_{g\in\mathcal{G}_{+,s}} |\hat{\mu}_g^2 - \mu_g^2| > \epsilon) &\leq G\exp(-\frac{1}{2}\frac{\epsilon^2/4}{4\bar{\rho}^2\bar{c}\frac{N}{G}\sigma_v^2\bar{k}}) + G\exp(-\frac{1}{2}\underline{\rho}^2\underline{kc}\frac{N}{G}) \\
&= \exp(\log G - \frac{C_1^2\log G}{32\bar{\rho}^2\bar{c}\sigma_v^2\bar{k}}) + \exp(\log G - \frac{1}{2\sigma_v^2}\underline{\rho}^2\underline{kc}\frac{N}{G}) \to 0.
\end{aligned}$$

[5]For any random variable $W \sim N(\mu, \sigma^2)$ $P(|W-\mu| \geq \sigma x) \leq 2e^{-x^2/2}$ for all $x \geq 0$.

Therefore, together with the result in Lemma A5 that with probability approaching one $\min_{g\in G_{+,s}} \mu_g^2$ is of order at least $\frac{N}{G}$, we know that with probability approaching one

$$\frac{\min_{g\in\mathcal{G}_{+,s}} \hat{\mu}_g^2}{\sqrt{\frac{N}{G}\log G}} \geq \frac{\min_{g\in\mathcal{G}_{+,s}} \mu_g^2}{\sqrt{\frac{N}{G}\log G}} - \frac{\max_{g\in\mathcal{G}_{+,s}} |\hat{\mu}_g^2 - \mu_g^2|}{\sqrt{\frac{N}{G}\log G}}$$

diverges to $+\infty$ if $(G\log G)/N \to 0$. Therefore take $\kappa_{G,N} = O\left(\sqrt{\frac{N}{G}\log G}\right)$, we have that with probability approaching one $\min_{g\in\mathcal{G}_{+,s}} \hat{\mu}_g^2/\kappa_{G,N}$ diverges as well and (i) is satisfied.

To prove ii), note

$$\max_{g\in\mathcal{G}_{+,s}^c} \hat{\mu}_g^2 = \max_{g\in\mathcal{G}_{+,s}^c} \{\mu_g^2 + \hat{\mu}_g^2 - \mu_g^2\} \leq \max_{g\in\mathcal{G}_{+,s}^c} \mu_g^2 + \max_{g\in\mathcal{G}_{+,s}^c} |\hat{\mu}_g^2 - \mu_g^2|.$$

For the second term on the right hand side, we have

$$\begin{aligned}
P(\max_{g\in\mathcal{G}_{+,s}^c} |\hat{\mu}_g^2 - \mu_g^2| > \eta) &\leq G \max_{g\in\mathcal{G}_{+,s}^c} P\left(|2\rho_g Z_g' v_g| + \frac{(Z_g' v_g)^2}{Z_g' Z_g} > \eta\right) \\
&\leq G \max_{g\in\mathcal{G}_{+,w}} P\Big(|2\rho_g Z_g' v_g| > \eta/2\Big) + G \max_{g\in\mathcal{G}_{+,s}^c} P\Big(\frac{(Z_g' v_g)^2}{Z_g' Z_g} > \eta/2\Big) \\
&\leq G \max_{g\in\mathcal{G}_{+,w}} 2\exp\left(-\frac{1}{2}\frac{\eta^2/4}{4a_g^2\sigma_v^2 k_g}\right) + G \max_{g\in\mathcal{G}_{+,s}^c} 2\exp(-\frac{1}{2}\frac{\eta/2}{\sigma_v^2}).
\end{aligned}$$

Let $\eta = C_2 \log G$ with any $C_2 > 4\sigma_v^2$, we know that $P(\sup_{g\in\mathcal{G}_{+,s}^c} |\hat{\mu}_g^2 - \mu_g^2| > \eta) \to 0$.

Therefore, together with the result in Lemma A5 that with probability approaching one $\max_{g\in G_{+,w}} \mu_g^2$ is bounded with probability approaching one. Then we know that for any $t > 1$,

$$\frac{\max_{g\in\mathcal{G}_{+,s}^c} \hat{\mu}_g^2}{(\log G)^t} \leq \frac{\max_{g\in\mathcal{G}_{+,s}^c} \mu_g^2}{(\log G)^t} + \frac{\max_{g\in\mathcal{G}_{+,s}^c} |\hat{\mu}_g^2 - \mu_g^2|}{(\log G)^t} = o_p(1).$$

Then (ii) is satisfied as long as $\log G/\kappa_{G,N} = o(1)$. For example the tuning sequence $\kappa_{G,N}$ can be of order $(\log G)^t$ for any $t > 1$ and at most $\sqrt{\frac{N}{G}\log G}$, which is well-defined since $G\log G/N \to 0$.

Let $\hat{\delta} = \hat{\mu}_{(\hat{K})}/\sqrt{\kappa_{G,N}}$. The above analysis has shown that with probability approaching one $\min_{g\in\mathcal{G}_{+,s}} \hat{\mu}_g > \max_{g\in\mathcal{G}_{+,s}^c} \hat{\mu}_g$ and $\hat{K} = G_{+,s}$. That is, all strong groups are selected and all other groups are not selected. Therefore, $\hat{\delta} \asymp \sqrt{N/(G\cdot\kappa_{G,N})}$ which meets Assumption 2.

Plug in $\hat{\delta}$ into $L(\hat{\delta}) = H^2 S_{sssel,int}(\hat{\delta})$ in Theorem 2, it is easy to see that $L(\hat{\delta}) = L^* + o_p(G/N)$. Then as $G, N \to \infty$ and $G \log G/N \to 0$,

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

□

Appendix C: Additional Simulation Results

Table A1: Standard Deviation 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	75.655	0.240	0.325	76.856	0.222	0.318	161.254	0.165	0.189	625.875	0.159	0.187
n=500	91.204	0.193	0.238	46.181	0.183	0.236	0.270	0.123	0.133	0.278	0.121	0.133
n=1000	26.381	0.146	0.167	32.785	0.141	0.165	0.175	0.089	0.094	0.177	0.088	0.093
	$G = 40$											
n=250	26.381	0.112	0.150	32.785	0.103	0.144	0.175	0.079	0.090	0.177	0.076	0.088
n=500	7.259	0.091	0.112	10.335	0.086	0.110	0.121	0.060	0.065	0.122	0.058	0.065
n=1000	0.291	0.071	0.083	0.300	0.068	0.081	0.084	0.044	0.047	0.085	0.043	0.046
	$G = 100$											
n=250	0.913	0.070	0.093	1.138	0.064	0.089	0.107	0.050	0.056	0.108	0.047	0.055
n=500	0.243	0.057	0.070	0.247	0.054	0.068	0.075	0.038	0.041	0.075	0.037	0.040
n=1000	0.166	0.044	0.052	0.168	0.043	0.051	0.053	0.028	0.029	0.053	0.027	0.029
	$G = 200$											
n=250	0.243	0.049	0.065	0.247	0.045	0.063	0.075	0.035	0.040	0.075	0.033	0.039
n=500	0.166	0.040	0.049	0.168	0.038	0.048	0.053	0.027	0.029	0.053	0.026	0.029
n=1000	0.115	0.031	0.037	0.116	0.030	0.036	0.037	0.020	0.021	0.037	0.019	0.021

Note: The table reports the standard deviation of the different estimators among 1000 simulations. The data generating process setups are the same as those used in Table 1.

Table A2: Finite-sample Bias 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	-0.192	0.123	0.038	-0.206	0.243	0.079	0.179	0.056	0.020	0.746	0.110	0.039
n=500	0.004	0.083	0.027	-0.104	0.156	0.046	-0.027	0.033	0.009	-0.048	0.061	0.014
n=1000	-0.484	0.049	0.026	-0.654	0.091	0.045	-0.015	0.018	0.009	-0.021	0.033	0.014
	$G = 40$											
n=250	-0.484	0.124	0.060	-0.654	0.247	0.120	-0.015	0.061	0.026	-0.021	0.121	0.053
n=500	-0.169	0.084	0.043	-0.318	0.165	0.080	0.003	0.036	0.014	-0.004	0.070	0.024
n=1000	-0.026	0.045	0.025	-0.047	0.094	0.054	-0.003	0.017	0.006	-0.005	0.035	0.014
	$G = 100$											
n=250	-0.046	0.123	0.063	-0.074	0.248	0.127	-0.008	0.061	0.026	-0.010	0.123	0.053
n=500	-0.011	0.079	0.040	-0.024	0.163	0.081	-0.001	0.034	0.013	-0.002	0.070	0.026
n=1000	-0.013	0.050	0.030	-0.020	0.100	0.061	-0.002	0.018	0.008	-0.003	0.038	0.017
	$G = 200$											
n=250	-0.011	0.124	0.064	-0.024	0.249	0.128	-0.001	0.062	0.028	-0.002	0.124	0.055
n=500	-0.013	0.084	0.043	-0.020	0.167	0.085	-0.002	0.035	0.012	-0.003	0.071	0.025
n=1000	-0.003	0.050	0.031	-0.008	0.100	0.063	-0.000	0.020	0.009	-0.001	0.038	0.018

Note: The table reports the finite-sample bias of the different estimators among 1000 simulations. The data generating process setups are the same as those used in Table 1.

References

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

Acemoglu, D., A. Finkelstein, and M. J. Notowidigdo (2013): "Income and Health Spending: Evidence from Oil Price Shocks," *Review of Economics and Statistics*, 95(4), 1079–1095.

Acemoglu, D., S. Johnson, J. Robinson, and P. Yared (2008): "Income and Democracy," *American Economic Review*, 98(3), 808–842.

Altmejd, A., A. Barrios-Fernández, M. Drlje, J. Goodman, M. Hurwitz, D. Kovac, C. Mulhern, C. Neilson, and J. Smith (2021): "O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries," *The Quarterly Journal of Economics*, 136(3), 1831–1886.

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

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14(1), 57–67.

Angrist, J. D., and J. S. Pischke (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., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80(6), 2369–2429.

Berkowitz, D., M. Caner, and Y. Fang (2008): "Are Nearly Exogenous Instruments Reliable?," *Economics Letters*, 101(1), 20–23.

Black, D., K. Daniel, and S. Sanders (2002): "The Impact of Economic Conditions on Participation in Disability Programs: Evidence from the Coal Boom and Bust," *American Economic Review*, 92(1), 27–50.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Card, D., F. Devicienti, and A. Maida (2014): "Rent-sharing, Holdup, and Wages: Evidence from Matched Panel Data," *Review of Economic Studies*, 81(1), 84–111.

Cervellati, M., F. Jung, U. Sunde, and T. Vischer (2014): "Income and Democracy: Comment," *American Economic Review*, 104(2), 707–719.

Charles, K., and M. Stephens, Jr. (2013): "Employment, Wages, and Voter Turnout," *American Economic Journal: Applied Economics*, 5(4), 111–43.

Charles, K. K., Y. Li, and M. Stephens, Jr. (2018): "Disability Benefit Take-up and Local Labor Market Conditions," *Review of Economics and Statistics*, 100(3), 416–423.

Cheng, X., Z. Liao, and R. Shi (2019): "On Uniform Asymptotic Risk of Averaging GMM Estimators," *Quantitative Economics*, 10(3), 931–979.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017): "Double/debiased/neyman Machine Learning of Treatment Effects," *American Economic Review*, 107(5), 261–265.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21(1), 1–68.

Coussens, S., and J. Spiess (2021): "Improving Inference from Simple Instruments through Compliance Estimation," *arXiv preprint arXiv:2108.03726*.

Currie, J., and E. Moretti (2003): "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings," *The Quarterly journal of economics*, 118(4), 1495–1532.

Deryugina, T., G. Heutel, N. H. Miller, D. Molitor, and J. Reif (2019): "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction," *American Economic Review*, 109(12), 4178–4219.

Dix-Carneiro, R., and B. K. Kovak (2017): "Trade Liberalization and Regional Dynamics," *American Economic Review*, 107(10), 2908–2046.

Donald, S. G., and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–1191.

Fredriksson, P., B. Ockert, and H. Oosterbeek. (2013): "Long-term Effects of Class Size," *The Quarterly journal of economics*, 128(1), 249–285.

Genovese, C., and L. Wasserman (2002): "Operating Characteristic and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society, B*, 64, 499–517.

Guggenberger, P. (2012): "On The Asymptotic Size Distortion of Tests When Instruments Locally Violate the Exogeneity Assumption," *Econometric Theory*, 28(2), 387–421.

Hahn, J., J. Hausman, and G. Kuersteiner (2004): "Estimation with Weak Instruments: Accuracy of Higher-Order Bias and MSE Approximations," *The Econometrics Journal*, 7(1), 272–306.

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

Jackson, C. K., R. C. Johnson, and C. Persico (2016): "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms," *The Quarterly Journal of Economics*, 131(1), 157–218.

Johnson, M. S. (2020): "Regulation by Shaming: Deterrence Effects of PublicizingViolations of Workplace Safety and Health Laws," *The American Economic Review*, 110(6), 1866–1904.

Kolesár, M. (2013): "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity," *Unpublished Working Paper*.

——— (2018): "Minimum Distance Approach to Inference with Many Instruments," *Journal of Econometrics*, 204(1), 86–100.

Leeb, H., and B. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21(1), 21–59.

Lleras-Muney, A. (2002): "Were State Laws on Compulsory Education Effective? An analysis from 1915 to 1939," *Journal of Law and Economics*, 45(2), 401–435.

Lleras-Muney, A. (2005): "The Relationship Between Education and Adult Mortality in the United States," *The Review of Economic Studies*, 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," *Journal of Econometrics*, 151(1), 1–16.

Oreopoulos, P. (2006): "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter," *American Economic Review*, 96(1), 152–175.

Pascali, L. (2017): "The Wind of Change: Maritime Technology, Trade, and Economic Eevelopment," *American Economic Review*, 107(9), 2821–2054.

Sølvsten, M. (2020): "Robust Estimation with Many Instruments," *Journal of Econometrics*, 214(2), 495–512.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

Stephens, M., and D.-Y. Yang (2014): "Compulsory Education and the Benefits of Schooling," *American Economic Review*, 104(6), 1777–1792.

Stock, J., and M. Yogo (2005): "Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. Andrews, and J. Stock, chap. 6, pp. 109–120. Cambridge University Press.

Wager, S., and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.