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

While most of the methodological literature on instrumental variable methods assumes homogeneity in the first-stage parameters, empirical applications of instrumental variable estimators often involve settings where the strength of the instruments varies depending on the composition of the sample. In this type of context, applied researchers have long recognized that restricting the analysis to subsamples with a strong first stage may help alleviate problems caused by weak instruments. For example, in the literature on return to compulsory schooling (e.g., Oreopoulos, 2006; Lleras-Muney, 2005) researchers have used variations in compulsory schooling laws to instrument for educational attainment. In this context, data suggest that Black students are weakly affected by changes in compulsory schooling laws and, as a result, instrumental variable estimation is often applied to samples of White students only.\footnote{The footnote 44 (page 209) of Lleras-Muney (2005) justifies the exclusion of Blacks, “Lleras-Muney (2002) shows, for example, that the laws affected Whites but not Blacks.”} Similarly, Cervellati, Jung, Sunde, and Vischer (2014) find 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 but much stronger for a sample of former colonies. Motivated by this finding, Cervellati, Jung, Sunde, and Vischer (2014) restrict the sample to former colonies and find a negative and significant effect of income on democracy. However, as we formally show later, sample selection based on the first stage correlation between the instrument and the endogenous variable produces invalid estimates. It tends to generate overly large biases of second-stage instrumental variable (IV) estimators, and overly large second-stage $t$-statistics under the null in significance tests.

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 instrumental variable 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. In practice, instrumental variable 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. The estimator is first-order efficient, yet may suffer from misleading inference in small sample due to many IV bias. 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 instrumental variable estimators as the criteria to form the decision rule of first stage multiple testing. 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 and selection from second-stage estimation to reduce asymptotic bias.

We argue that the proposed adaptive procedure is preferred to the naive direct selection method as well as other existing methods adopted by practitioners such as pooled 2SLS or fully interacted 2SLS. We find that another estimator that performs well in an IV model with first stage heterogeneity is the limited information maximum likelihood (LIML) estimator that uses a full set of interacted instruments to account for first-stage heterogeneity, while the Monte Carlo simulations show that the proposed adaptive estimator outperforms fully-interacted LIML for data generating processes (DGPs) that have a small proportion of strong groups.

Our proposed estimation procedure is related to Donald and Newey (2001), Okui (2009), Cheng, Liao, and Shi (2015), and others who use higher-order asymptotic expansion as in Nagar (1958) to choose instruments. The instrument selection idea is also related to the machine learning approach for IV regression such as Belloni, Chen, Chernozhukov, and Hansen (2012) and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). The key difference between the MSE expansion literature that we follow and the machine learning literature is that the IV selection criteria of the former is based on the asymptotic MSE of the second-stage estimator while that of the latter is based on first-stage fitting. Our
paper is also related to a recent work by Coussens (2019) who studies treatment evaluation with endogeneity and proposes to reweigh 2SLS by individual compliance propensity estimated through cross-validated causal forest (c.f. Wager and Athey, 2018). Although both papers explore the possibility of utilizing first-stage heterogeneity to improve the precision of second-stage IV estimation, different model-setups and methods are considered. Motivated by our empirical applications, we explicitly model subsamples that might lack first-stage identification and propose to adaptively deselect such groups based on the criteria of minimizing higher-order terms in the asymptotic MSE formula of the second-stage estimator.

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 first discuss the inconsistency of the naive direct selection approach often used in applied work. Then, 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. We confirm the MSE improvement of the proposed adaptive estimator in simulation studies in Section 3. We conclude the article with empirical applications to the compulsory schooling data of Oreopoulos (2006) and the voter turnout data of Charles and Stephens (2013).

2. Theory

2.1. Model Set-up

As discussed in the introduction, in applied settings it is often the case that the correlation between an endogenous variable and an instrument is heterogeneous across different population subgroups. Consider a simultaneous equation model with a heterogeneous first stage where the instrument is strong for some population subgroups, weak for some other subgroups, and uncorrelated with the endogenous variable for the rest. The model is a natural specification of a variety of economic applications. For example, in the return to compulsory schooling literature, 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 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, \( \tilde{Z} \). The model has heterogeneous first stage coefficients across groups. Assume
that for each individual \( i \) in group \( g \), we have

\[
Y_{ig} = \beta W_{ig} + X_{ig} \theta + u_{ig}, \\
W_{ig} = \rho_g \tilde{Z}_{ig} + X_{ig} \gamma + v_{ig},
\]

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 a non-zero correlation between \( u_{ig} \) and \( v_{ig} \). In empirical research,
groups could be determined by observables like geographic regions, ethnic groups, etc.

In matrix form, we can re-write (1) as

\[
Y_g = \beta W_g + X_g \theta + u_g, \\
W_g = \tilde{Z}_g \rho_g + X_g \gamma + v_g
\]

where \( Y_g, W_g \) and \( \tilde{Z}_g \) are vectors of length \( n_g \) and \( X_g \) is matrix of dimension \( n_g \times d \). Group-
wise transform the instrument and re-write the first stage equation to obtain

\[
W_g = Z_g \rho_g + X_g \omega_g + v_g
\]

where \( Z_g = M_{X_g} \tilde{Z}_g \) 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 conditional on group,
   \((\tilde{Z}_{ig}, X_{ig}, Y_{ig}, W_{ig})\) are i.i.d. across \( i = 1, \ldots, n_g \). There exist positive and finite \( c \) and
   \( \bar{c} \) such that \( \frac{c}{G} \leq n_g \leq \frac{\bar{c}}{G} \) for all \( g = 1, 2, \ldots, G \) as \( N, G \to \infty \).
2. One-sided First Stage Relationship: There exist positive and finite $\rho$ and $\bar{\rho}$ such that $\rho \leq a_g < \bar{\rho}$ for all $g = 1, \ldots, G$. Let groups with irrelevant IV be defined as $G_0 = \{g : \rho_g = 0\}$, groups with strong IV be defined as $G_{+,s} = \{g : \rho_g = a_g\}$, and groups with weak IV be defined as $G_{+,w} = \{g : \rho_g = a_g/\sqrt{n_g}\}$. Denote $G_0 = |G_0|$, $G_{+,s} = |G_{+,s}|$ and $G_{+,w} = |G_{+,w}|$. Let $G_+ = G_{+,w} \cup G_{+,s}$ and $G_+ = G_{+,w} + G_{+,s}$.

3. Moment Conditions: Let $k_g = \lim_{n_g \to \infty} E[\tilde{Z}_g'Z_g/n_g]$. There exist positive and finite $k$ and $\bar{k}$ such that $k \leq k_g \leq \bar{k}$ for all $g = 1, \ldots, G$. In addition, there exists a positive and finite constant that bounds $E[\tilde{Z}_{ig}^8]$ as well as $E[X_{ig}^8]$ uniformly across all $g = 1, \ldots, G$.

4. Error Terms: For all $g = 1, \ldots, G$, $(u_{ig}, v_{ig})(\tilde{Z}_{ig}, \tilde{X}_{ig})$ have a common distribution with mean 0 and non-singular variance-covariance matrix $(\sigma_u^2, \sigma_{uv}; \sigma_{uv} \sigma_v^2)$. In addition, there exists a positive and finite constant that bounds $E[\tilde{v}_{ig}^8]$ across all $g = 1, \ldots, G$.

5. Nontrivial Presence of Strong Groups: When $G$ is fixed, $G_{+,s} > 0$. When $G, N \to \infty$, $G_{+,s}/G \to b \geq \bar{b} > 0$.

Assumption 1.1 allows unbalanced group size but requires that all groups have sample sizes of the same order for both the fixed $G$ 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 for conceptual and notational simplicity and is a natural assumption in our two empirical applications where we have compulsory schooling law as an instrument to years of education, and energy supply shocks as an instrument to local economic conditions. Without loss of generality, we assume that first-stage effects are positive or zero.\(^2\) When the first stage coefficient is zero, the instrument is irrelevant. When the first stage coefficient is of order $O(1/\sqrt{n_g})$, the instrument is weak. If $G_{+,w} = 0$, there are only zero and positive first-stage coefficients so the positive first stage coefficients are well separated from zero. Assumption 1.3 requires that the instrument has nontrivial variations in each group. In applications, groups with zero variations in the instrument need to be dropped out in advance, which is

\(^2\)We expect results of the paper to carry through without the sign restriction in Assumption 1.2. See brief discussions in the proof of Theorem 1.
in line with the LATE interpretation that 2SLS is only measuring an average second-stage effect among individuals who respond first-stage to the variation in the instrument. 

Assumption 1.4 imposes exclusion restrictions and homoskedasticity on the distributions of the error terms. These assumptions are commonly adopted in the literature. Assumption 1.5 is required for the identification of $\beta$. Similar assumptions of strong identification are often employed in the IV literature. For example, Okui (2009) and Cheng, Liao, and Shi (2015) 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. To avoid complications in the interpretation of the estimated parameter (see, e.g. Angrist and Imbens, 1995 and Abadie, 2003), we formally assume that $\beta$ is constant.

### 2.2. Existing Methods

Let $Y, W, X, Z, u$ and $v$ be vectors or matrices with row size $N$ that stack all group vectors $Y_g, W_g, X_g, Z_g, u_g$ and $v_g$ together. In empirical studies, two estimation strategies are often used for the set-up described above. The pooled 2SLS or IV estimator ignores the first stage effect heterogeneity and runs an IV regression on the pooled data. The fully-interacted 2SLS estimator incorporates potential heterogeneity in the first-stage parameter, running a 2SLS with the instrument interacted with group indicators. This estimator may suffer from many-IV bias especially when the number of groups is large. For any full-rank matrix $A$, let $P_A = A(A'A)^{-1}A'$ and $M_A = I - P_A$. Let $D$ is a block diagonal matrix whose block diagonal elements are $Z_1, \ldots, Z_G$, and $\hat{\rho}_g = (Z'_gZ_g)^{-1}(Z'_gW_g)$ is the group-wise first-stage estimator for $\rho_g$. Let $\hat{\beta}_{pool}$ and $\hat{\beta}_{int}$ be the pooled and fully-interacted IV estimators, respectively. Since $Z'_gX_g = 0$ for all $g = 1, \ldots, G$, we have that

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

\[
\hat{\beta}_{int} = (W'P_DW)^{-1}W'PDY = \left(\sum_{g=1}^{G} \hat{\rho}_g Z'_g W_g\right)^{-1} \sum_{g=1}^{G} \hat{\rho}_g Z'_g Y_g
\]
The definition of $\hat{\beta}_{\text{int}}$ in (3) implies that the fully-interacted estimator is equivalent to a weighted IV estimator using first stage slope estimators in each group to reweight the instrumental variable.

Under Assumption 1, the pooled IV estimator $\hat{\beta}_{\text{pool}}$ is consistent and

$$\sqrt{N} \left( \hat{\beta}_{\text{pool}} - \beta \right) \Rightarrow N \left( 0, \sigma_u^2 \bar{H}_p^{-1} \right),$$

where $\bar{H}_p = \lim_{N,G \to \infty} E \left[ \sum_{g=1}^{G} \rho_g Z'_g Z_g / N \right] / E \left[ \sum_{g=1}^{G} Z'_g Z_g / N \right]$. Similarly, provided that $G^2 / N \to 0$, the fully-interacted estimator $\hat{\beta}_{\text{int}}$ is also consistent and

$$\sqrt{N} \left( \hat{\beta}_{\text{int}} - \beta \right) \Rightarrow N \left( 0, \sigma_u^2 \bar{H}^{-1} \right),$$

where $\bar{H} = \lim_{N,G \to \infty} E \left[ \sum_{g=1}^{G} \rho_g^2 Z'_g Z_g / N \right]$. Note that by the Cauchy-Schwarz inequality we know

$$\left( \sum_{g=1}^{G} \rho_g \sqrt{E \left[ Z'_g Z_g \right]} \cdot \sqrt{E \left[ Z'_g Z_g \right]} \right)^2 \leq \sum_{g=1}^{G} \rho_g^2 E \left[ Z'_g Z_g \right] \cdot \sum_{g=1}^{G} E \left[ Z'_g Z_g \right]$$

and therefore $\bar{H} \geq \bar{H}_p$ unless the first stage is homogeneous or $\rho_g = \rho$ for all $g = 1, \ldots, G$. In fact, the fully-interacted estimator has the same asymptotic limit as the infeasibly weighted IV estimator

$$\hat{\beta}_{\text{inf}} = \left( \sum_{g=1}^{G} \rho_g Z'_g W_g \right)^{-1} \sum_{g=1}^{G} \rho_g Z'_g Y_g \quad (4)$$

where the instrument is reweighted using oracle values for $\rho_g$. Such an infeasible estimator is efficient under homoskedasticity, as in Assumption 1. If the homoskedasticity assumption is violated, efficient estimation of $\beta$ would involve a GLS-type of reweighting based on estimated variance of the second-stage error term as well.

The growth condition $G^2 / N \to 0$ for the fully-interacted estimator is required to guarantee that the asymptotic bias of the full interacted 2SLS estimator vanishes when $N, G \to \infty$. Even when this condition is satisfied, the estimator may suffer from “many IV bias” in small samples (Staiger and Stock, 1997; Stock and Yogo, 2005). In such cases, empirical researchers might choose the pooled estimator in the hope of having correct size control in significance tests, or correct coverage for confidence intervals. However, if the first stage effect is highly heterogeneous and the instrumental variable is relevant only for some groups, pooling all groups together would result in a substantially inefficient estimator.
As discussed in the introduction, a selective IV regression approach is often used by applied researchers aiming to obtain a strong first stage. This approach consists of running an IV regression using only the groups selected by testing $H_{0,g} : \rho_g = 0$ against the alternative $H_{a,g} : \rho_g > 0$, $g = 1, \ldots, G$, with some pre-determined significance level $\alpha_{FS}$. Let $t_g$ be the $t$-statistic for group $g$ and $c_{g,\alpha_{FS}}$ be the $(1 - \alpha_{FS})$ quantile of Student-t with $n_g - d - 1$ degrees of freedom. Assuming that at least one group is selected, the resulting estimator is

$$\hat{\beta}_{sel,pool} = \left( \sum_{g=1}^{G} Z_g'W_g 1(t_g > c_{g,\alpha_{FS}}) \right)^{-1} \sum_{g=1}^{G} Z_g'Y_g 1(t_g > c_{g,\alpha_{FS}})$$ (5)

which we will refer to as the select-and-pool estimator. The next theorem shows that for this estimator, the exclusion restriction is in fact violated at a rate that invalidates the conventional inference.

**Theorem 1.** Let $c_{g,\alpha_{FS}} \equiv P_{\nu_g-d-1}^{-1}(1 - \alpha_{FS})$ be the $(1 - \alpha_{FS})$-quantile of a Student-t distribution with degrees of freedom $n_g - d - 1$. Suppose Assumption 1 holds, and let $\sigma_{uv} \neq 0$ and $0 \leq \alpha_{FS} < 1/2$. Then

1. if $G$ is fixed and $G_{+,s} < G$, there exists a positive constant $a$ such that

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

2. if $G, N \to \infty$ and $G_{+,s}/G \to b \leq \bar{b} < 1$, there exists a positive constant $a$ such that

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

Theorem 1 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 mean that the concerned IV estimator is inconsistent. Nor does it imply that the estimator has incorrect inference at the limit. Guggenberger (2012), however, shows that if the exclusion restriction of the instrument is violated locally at a rate $1/\sqrt{N}$, many commonly used inference methods such as the Wald test, the Anderson Rubin or the conditional likelihood ratio tests have asymptotic size distortion. Theorem 1 shows that if pointwise testing is employed, the exclusion restriction of the select-and-pool method is violated at a rate no smaller than $1/\sqrt{N}$ when $G$ is fixed, as long as not all groups are strong. When $G$ grows together with $N$, the exclusion restriction is violated at a rate higher than $1/\sqrt{N}$ as long as groups with weak or irrelevant IV have a non-vanishing proportion. Under such circumstances, the type I error of conventional inference methods based on the select-and-pool estimator goes to one. For the second case of increasing $G$, similar to Theorem 1, one could also show that if the Bonferroni type multiple testing method is adopted such that groupwise cutoff is equal to $P_{n \sigma - d - 1}^{-1}(1 - \alpha_{FS}/G)$, the exclusion restriction of the select-and-pool procedure is violated at a rate no smaller than $1/\sqrt{NG}$. But if $G$ grows much slower than the average sample size $N/G$, the $1/\sqrt{NG}$ rate could still be close to the local violation rate of $1/\sqrt{N}$, implying that the estimator could suffer from significant finite sample size distortion with traditional inference methods. More importantly, as we will discuss in the next section, the select-and-pool estimator, or more generally, pooled IV estimation, is not efficient under first stage heterogeneity.
alleviated with the increase of sample size. When over-rejects more severely when the number of groups grows, and the size distortion is not level test. As predicted by Theorem 1, the test based on the select-and-pool estimator used for the simulations are specified in the footnote of the table. Each number in the table fully interacted estimator increases with the number of groups, the degree of endogeneity, from the "many IV bias". Similar to the select-and-pool estimator, the size distortion of the inefficient, as illustrated in Table A1 in the Appendix which reports the variability of each estimators under the same DGPs. The pooled estimator controls size well but is highly inefficient, as illustrated in Table A1 in the Appendix which reports the variability of each estimator reported in Table 1. As we discussed earlier, the fully interacted estimator suffers from the "many IV bias". Similar to the select-and-pool estimator, the size distortion of the fully interacted estimator increases with the number of groups, the degree of endogeneity, and the proportion of groups with irrelevant instruments. In some simulations (e.g., $G = 50$

Table 1: Size-Distortion of Existing Estimators

<table>
<thead>
<tr>
<th>$\rho_{uv}$</th>
<th>$G_{+,s}/G = 0.1$</th>
<th>$G_{+,s}/G = 0.2$</th>
<th>$G_{+,s}/G = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{pool}$</td>
<td>$\beta_{int}$</td>
<td>$\beta_{sel,pool}$</td>
<td>$\beta_{sel,int}$</td>
</tr>
<tr>
<td>$\rho_{uv} = 0.6, G = 10$</td>
<td>$n = 500$</td>
<td>0.043</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td>$n = 1000$</td>
<td>0.041</td>
<td>0.078</td>
</tr>
<tr>
<td></td>
<td>$n = 500$</td>
<td>0.060</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>$n = 1000$</td>
<td>0.052</td>
<td>0.088</td>
</tr>
<tr>
<td>$\rho_{uv} = 0.8, G = 10$</td>
<td>$n = 500$</td>
<td>0.047</td>
<td>0.235</td>
</tr>
<tr>
<td></td>
<td>$n = 1000$</td>
<td>0.041</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>$n = 500$</td>
<td>0.052</td>
<td>0.361</td>
</tr>
<tr>
<td></td>
<td>$n = 1000$</td>
<td>0.053</td>
<td>0.196</td>
</tr>
<tr>
<td>$\rho_{uv} = 0.6, G = 100$</td>
<td>$n = 500$</td>
<td>0.041</td>
<td>0.374</td>
</tr>
<tr>
<td></td>
<td>$n = 1000$</td>
<td>0.060</td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td>$n = 500$</td>
<td>0.053</td>
<td>0.605</td>
</tr>
<tr>
<td></td>
<td>$n = 1000$</td>
<td>0.057</td>
<td>0.372</td>
</tr>
</tbody>
</table>

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 $X_{ig} \sim \text{i.i.d.} \ N(0, 1)$, $(u_{ig}, v_{ig}) \sim N((0, 0), (\rho_{uv}; \rho_{uv} = 1))$, $W_{ig} = \rho_g \tilde{Z}_{ig} + X_{ig} + v_{ig}$, $Y_{ig} = \beta W_{ig} + X_{ig} + u_{ig}$ for $i = 1, 2, \ldots, n$, where $\beta = 0$, $\rho_g = 0.5$ for $g = 1, \ldots, G_{+,s}$ and $\rho_g = 0$ for $g > G_{+,s}$.}

Table 1 illustrates the over-rejection problem of the select-and-pool estimator. DGPs used for the simulations are specified in the footnote of the table. Each number in the table reports the rejection proportion among 1000 simulations for a nominal 5 percent significance level test. 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 16 percent. When $G = 100$, the rejection rate becomes as high as around 50 percent. The over-rejection problem also gets worse with increased model endogeneity and zero group proportion.

Table 1 also reports the small sample performance of the pooled and the fully interacted estimators under the same DGPs. The pooled estimator controls size well but is highly inefficient, as illustrated in Table A1 in the Appendix which reports the variability of each estimator reported in Table 1. As we discussed earlier, the fully interacted estimator suffers from the “many IV bias”. Similar to the select-and-pool estimator, the size distortion of the fully interacted estimator increases with the number of groups, the degree of endogeneity, and the proportion of groups with irrelevant instruments. In some simulations (e.g., $G = 50$
or 100, \( n = 500 \)), the fully interacted estimator has worse size distortion than the select-and-pool estimator. However, since “many IV bias” is a finite sample problem, its size distortion will improve as the sample size grows. Table A2 in the Appendix provides a detailed comparison of bias of different estimators.

Table 1 also reports the small sample performance of a select-and-interact estimator, which is a variation of the select-and-pool estimator. This estimator interacts the instrument with group dummies using only data for groups that pass the first-stage selection test with a five percent significance level. Interestingly, this estimator has much better size-distortion behavior than both the fully-interacted and the select-and-pool estimator. This motivates the study the estimator defined in the next section.

2.3. The Select-and-Interact Estimators

First, we define the select-and-interact estimator \( \hat{\beta}_{sel,int}(\delta) \) to be the same as the fully-interacted estimator defined in (3) except that only groups that pass some selection rule will be used in estimation. We consider the selection rule \( 1(\hat{\mu}_g > \delta) \) where \( \hat{\mu}_g = \hat{\rho}_g(Z'_gZ_g)^{1/2} = (Z'_gZ_g)^{-1/2}Z'_gW_g \) is the first-stage signal in group \( g \) motivated by the form of the asymptotic variance of the infeasible fully-interacted estimator \( \hat{\beta}_{inf} \), and \( \delta \) is the selection cut-off. The estimator is hence defined as

\[
\hat{\beta}_{sel,int}(\delta) = \left( \sum_{g=1}^{G} \hat{\rho}_gZ'_gW_g1(\hat{\mu}_g > \delta) \right)^{-1} \sum_{g=1}^{G} \hat{\rho}_gZ'_gY_g1(\hat{\mu}_g > \delta).
\]

Under homoskedasticity, the ranking of \( \hat{\mu}_g \) among groups is asymptotically equivalent to the ranking of the \( t_g \) statistics defined in the last section. The infeasible version of the selection metric, \( \mu_g = \rho_g(Z'_gZ_g)^{1/2} \), indicates that both the first stage coefficient \( \rho_g \) and the amount of variation in the instrument are essential for determining the contribution that a specific group brings to the second-stage estimation. This can be seen in the asymptotic variance of the infeasibly weighted IV estimator defined in (4). It is reasonable to expect that the range of \( \delta \) can not be unrestricted, because if \( \delta \) is too large such that we discard groups with strong instruments, then the estimator suffers from efficiency loss. To guarantee a first-order efficient estimator, we restrict \( \delta \) to fall inside of the following range.
Assumption 2. (Range of $\delta$) The thresholding value $\delta \in \Delta = \left\{ \delta : \delta \leq C_\delta (N/G)^{1/2} \right\}$ for some constant $C_\delta < \sqrt[4]{k}c$.

The range of the $\delta$ in the assumption guarantees that as $N, G$ grows, all groups with strong first stage are selected such that the first-order efficiency bound of the estimator is reached asymptotically. More specifically, the upper bound of the range guarantees that the probability of committing type II error for the strong groups vanishes sufficiently fast. See Lemma A3 in the Appendix for details. The range, therefore, accommodates first stage testing procedures with a fixed nominal size. It also includes other testing procedures that adjust for the growing number of hypotheses in the first stage, including the Bonferroni correction or 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.

The following lemma gives the asymptotic property of the select-and-interact estimator.

Lemma 1. Under Assumption 1 and 2 and if additionally $G^2/N \to 0$ as $G, N \to \infty$, then

$$\sqrt{N}(\hat{\beta}_{\text{sel,int}}(\delta) - \beta) \Rightarrow N(0, \sigma_u^2 \bar{H}^{-1}).$$

Unlike the select-and-pool method discussed in the previous section, the select-and-interact estimator is consistent with valid asymptotic inference using conventional methods as long as the growth condition, $G^2/N \to 0$, is satisfied. The positive result may be unexpected. Intuitively, it comes from the fact that by interacting the instrument with indicators for the groups that survive the first stage selection, the estimator 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. It turns out that 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 $1/\sqrt{N}$. When $\delta = -\infty$, the select-and-interact estimator is equivalent to the fully-interacted estimator $\hat{\beta}_{\text{int}}$ defined in (3).

(2018) and others for conducting causal inference with machine learning methods, we also introduce split-sample select-and-interact estimator where the first-stage selection and the second stage estimation are separated with two subsamples obtained by random data splitting.

Randomly partition the data into two equal proportions within each groups and index the resulting two samples by $a$ and $b$. Denote random variables and matrices with superscript $a$ and $b$ for the corresponding subsamples. Let $n^a_g = \lceil n_g/2 \rceil$ and $N^a = \sum_{g=1}^G n^a_g$. Let $\hat{\rho}^a_g = ((Z^a_g)'Z^a_g)^{-1}(Z^a_g)'W^a_g$, $\hat{\mu}^a_g = ((Z^a_g)'Z^a_g)^{-1/2}(Z^a_g)'W^a_g$ and define similar terms for subsample $b$. Let

$$
\hat{\beta}^a(\delta) = \left( \sum_g \hat{\rho}^b_g(Z^a_g)'W^a_g 1(\hat{\mu}^b_g \geq \delta) \right)^{-1} \sum_g \hat{\rho}^b_g(Z^a_g)'Y^a_g 1(\hat{\mu}^b_g \geq \delta),
$$

$$
\hat{\beta}^b(\delta) = \left( \sum_g \hat{\rho}^a_g(Z^b_g)'W^b_g 1(\hat{\mu}^a_g \geq \delta) \right)^{-1} \sum_g \hat{\rho}^a_g(Z^b_g)'Y^b_g 1(\hat{\mu}^a_g \geq \delta), \quad \text{and}
$$

$$
\hat{\beta}_{ss,int}(\delta) = \left( \hat{\beta}^a(\delta) + \hat{\beta}^b(\delta) \right) / 2. \quad (7)
$$

Both $\hat{\beta}^a(\delta)$ and $\hat{\beta}^b(\delta)$ separate out first-stage selection and instrument reweighting from second-stage estimation using different subsamples. By averaging across $\hat{\beta}^a(\delta)$ and $\hat{\beta}^b(\delta)$, the repeated split-sample select-and-interact estimator, defined in (7), preserves efficiency as stated in the next lemma. When $\delta = -\infty$, the estimator also reduces to

$$
\hat{\beta}_{ss,int} = \left( \hat{\beta}^a_{int} + \hat{\beta}^b_{int} \right) / 2 \quad (8)
$$

where $\hat{\beta}^a_{int} = \left( \sum_g \hat{\rho}^b_g(Z^a_g)'W^a_g \right)^{-1} \sum_g \hat{\rho}^b_g(Z^a_g)'Y^a_g$ and similarly for $\hat{\beta}^b_{int}$. This is a repeated split-sample version of $\hat{\beta}_{int}$, defined earlier in (3).

**Lemma 2.** Under Assumption 1 and 2 and if additionally $G/N \rightarrow 0$ as $G, N \rightarrow \infty$, then

$$
\sqrt{N}(\hat{\beta}_{ss,int}(\delta) - \beta) \Rightarrow N(0, \sigma_u^2 \hat{H}^{-1}).
$$

Lemmas 1 and 2 imply that full sample and repeated split-sample select-and-interact estimators, defined respectively in (6) and (7) with different cut-off values $\delta \in \Delta$ are asymptotically equivalent. This equivalence result, however, will not materialize in finite samples.
In fact, the weaker growth condition between $G$ and $N$ required in Lemma 2 suggests that the higher-order bias and/or higher-order efficiency loss terms of the split-sample select-and-interact estimator might be smaller in order of magnitudes compared to the full sample select-and-interact estimator. The next theorem confirms this argument by formulating the asymptotic MSEs of the two estimators as a function of the value $\delta$. To keep the calculation tractable, we assume for the theorem 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

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

$$N(\hat{\beta}_{sel,int}(\delta) - \beta)^2 = \hat{Q}_{sel,int}(\delta) + \hat{r}_{sel,int}(\delta)$$

$$E[\hat{Q}_{sel,int}(\delta)|\tilde{Z},X] = \sigma_u^2/H + S_{sel,int}(\delta) + T_{sel,int}(\delta),$$

$$(\hat{r}_{sel,int}(\delta) + T_{sel,int}(\delta))/S_{sel,int}(\delta) = o_p(1),$$

where $H = \frac{1}{N} \sum_{g \in G} g \rho_{g}^2 Z_g' Z_g$ and

$$H^2 S_{sel,int}(\delta) = \left( \sigma_{uv} \sum_{g} \left( 1 - \Phi \left( \frac{\delta - \mu_g}{\sigma_v} \right) + \left( \frac{\delta - \mu_g}{\sigma_v} \right) \phi \left( \frac{\delta - \mu_g}{\sigma_v} \right) \right) + \frac{\sigma_{uv} \sum_{g} \mu_g}{\sigma_v} \phi \left( \frac{\delta - \mu_g}{\sigma_v} \right) \right)^2 / N.$$

2. if $G/N \rightarrow 0$ as $G,N \rightarrow \infty$, the asymptotic MSE of $\hat{\beta}_{sssel,int}$ can be decomposed as

$$N(\hat{\beta}_{sssel,int}(\delta) - \beta)^2 = \hat{Q}_{sssel,int}(\delta) + \hat{r}_{sssel,int}(\delta)$$

$$E[\hat{Q}_{sssel,int}(\delta)|\tilde{Z},X] = \sigma_u^2/H + S_{sssel,int}(\delta) + T_{sssel,int}(\delta)$$

$$(\hat{r}_{sssel,int}(\delta) + T_{sssel,int}(\delta))/S_{sssel,int}(\delta) = o_p(1)$$

14
where $H^2 S_{ssel,int}(\delta) = A_{ssel,int}(\delta) + B_{ssel,int}(\delta) + C_{ssel,int}(\delta)$ with

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

$$B_{ssel,int}(\delta) = \sigma_u^2 \sum_g \mu_g^2 \Phi \left( \frac{\delta - \mu_g / \sqrt{2}}{\sigma_v} \right) / N$$

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

For the full-sample select-and-interact estimator $\hat{\beta}_{ssel,int}$, the first-order term in its asymptotic MSE decomposition is the variance term $\sigma_u^2 / H$, 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 $\delta$ 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 $\sigma_{uv}^2 \left( \sum_g \left( 1 - \Phi \left( \frac{\delta - \mu_g / \sqrt{2}}{\sigma_v} \right) \right) \right)^2 / N$. The additional terms in $S_{ssel,int}(\delta)$ then represent extra estimation bias introduced from first-stage selection through the correlation between the error terms in the first and second stages.

The repeated split-sample estimator $\hat{\beta}_{ssel,int}$ has the same first-order MSE term as the full-sample select-and-interact estimator but its second-order leading term is only of order $G/N$. The term comes from two different sources. $A_{ssel,int}(\delta)$ and $B_{ssel,int}(\delta)$ contain two higher-order variance components and $C_{ssel,int}(\delta)$ is a higher-order bias term. $A_{ssel,int}(\delta)$ represents higher-order efficiency loss from using more instruments in IV regression. $B_{ssel,int}(\delta)$ represents higher-order efficiency loss from falsely excluding groups with relevant instruments. $C_{ssel,int}(\delta)$ is a bias term coming from first-stage selection and the combination of two split-sample estimators. The bias arises from the correlation between the error terms in the first stage of $\hat{\beta}^a(\delta)$ (or $\hat{\beta}^b(\delta)$) and the second stage of $\hat{\beta}^b(\delta)$ (or $\hat{\beta}^a(\delta)$). As $\delta$ increases, $A_{ssel,int}(\delta)$ and $C_{ssel,int}(\delta)$ decrease, while $B_{ssel,int}(\delta)$ increases. This leads us to propose a data-driven approach to choose $\delta$ that minimizes the asymptotic MSE of
2.4. Adaptive $\delta$ selection for optimal MSE

In this section, we show that we can choose the thresholding value $\delta$ adaptively to achieve the optimal asymptotic MSE derived in Theorem 2 when there is strong separation between groups. This is an intuitive result, as strong separation would allow the adaptive choice of $\delta$ to successfully distinguish the groups with effective IV from those without, thus minimizing the higher-order leading term in the asymptotic MSE formula. Corollary 1 first establishes the optimal level of the asymptotic MSE of the repeated split-sample select-and-interact estimator.

**Corollary 1.** Denote $L(\delta) \equiv H^2S_{\text{sssel,int}}(\delta)$. Under the assumptions of Theorem 2 and suppose $G_{+,w}/G \to 0$, then $\min_\delta L(\delta) = 2b\sigma_u^2\sigma_v^2(1+\rho_{uv})\frac{G}{N} + o_p(G/N)$, where $b = \lim_{G \to \infty} G_{+,s}/G$.

When the group with weak first stage coefficient vanishes, the minimum asymptotic MSE level is achieved when the thresholding value of $\delta$ separates out groups with and without first-stage identification. Therefore, the optimal constant in front of the factor $G/N$ depends on the proportion of effective groups, rather than the specific configurations of $\mu_g$ for $g \in G_{+,s}$.

The next theorem suggests an adaptive estimator for the optimal thresholding value $\delta$.

**Theorem 3.** Assume $G_{+,w} = 0$ and let $(\hat{\sigma}_u^2, \hat{\sigma}_v^2, \hat{\sigma}_{uv}^2)$ be some consistent estimators of $(\sigma_u^2, \sigma_v^2, \sigma_{uv}^2)$. Let

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

for $\hat{\mu}(1) \geq \hat{\mu}(2) \geq \hat{\mu}(G_{+,s}) > \hat{\mu}(G_{+,s}+1) = \ldots = \hat{\mu}(G) = 0$, and

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

where $\hat{\mu}(1) \geq \hat{\mu}(2) \geq \hat{\mu}(G)$ with $\hat{\mu}_g = \hat{\rho}_g\sqrt{Z_g'Z_g}$ and $\kappa_N = \mathcal{C} \sqrt{\frac{N\log G}{G}}$ for some positive constant $\mathcal{C} < \infty$. Finally, let $\hat{K} = \arg\min_K \hat{\mathcal{R}}(K)$, Under the assumptions of Theorem 2
and provided that $G \log G/N \to 0$, we have
\[
\frac{\mathcal{R}(\hat{K})}{\min L(\delta)} \overset{p}{\to} 1.
\]

Theorem 3 implies that we can pick the adaptive thresholding value $\hat{\delta}$ to be the $\hat{K}$-th order statistics of $\hat{\mu}$ such that the resulting $\hat{\beta}_{\text{ssel, int}}(\hat{\delta})$ achieves the minimum asymptotic MSE. The tuning parameter $\kappa_N$ is used as a wedge to separate the groups with strong first-stage signal from those with irrelevant instruments when the first stage parameter $\rho_g$ has to be replaced by its estimator. Intuitively, $\kappa_N$ is chosen to dominate all $\hat{\mu}_g$ terms in $\mathcal{G}_0$ groups and be dominated by all $\hat{\mu}_g$ terms in $\mathcal{G}_{+,s}$ groups such that $\hat{R}(\dot{\cdot})$ is minimized at a value that will include all strong groups but discard all zero groups in the limit. The scaling constant $C$ is set to 0.1 in the simulation section\textsuperscript{3} and two empirical examples based on simulation evidence.

When the set of weak groups $\mathcal{G}_{+,w}$ is not empty, the minimum asymptotic MSE is still of order $G/N$, but the optimal constant depends on the data generating process for $\mu_g$.\textsuperscript{4} 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 the MSE of $\hat{\beta}_{\text{ssel, int}}$ with an adaptive estimated optimal $\hat{\delta}$ to converge in probability to the optimal MSE in the presence of weak instruments.

3. Monte Carlo Simulations

In this section, we study the finite sample performance of the estimation procedures analyzed in this article. We use three data generating processes. Let $X_i, \tilde{Z}_i, v_i, e_i \sim i.i.d. N(0, 1)$, and $u_i = \rho_{u,v} v_i + \sqrt{1 - \rho_{u,v}^2} e_i$. 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 $\rho_g$ controls the relevance of instrument $Z$ in group $g$ and varies across DGPs. Define $\lambda_g = n_g \rho_g^2$ as the
\textsuperscript{3}We try different constant of $C$ in the Monte Carlo experiments and $C = 0.1$ seems to perform best for different data generating processes.
\textsuperscript{4}When $G_2/G \to b_2 > 0$, the minimum asymptotic MSE is of order $G/N$, but the constant in front of it depends on the distribution of $\mu_g$ for $g \in \mathcal{G}_{+,w}$, often in a very complicated fashion.
Figure 1 summarizes the distribution of $\rho_g$ for the three DGPs. We fix group size to $n_g = 500$ throughout. DGP 1 represents the case with well-separation among groups in terms of strength of first-stage signal. Out of $G$ groups, where $G$ varies from 40 to 400 in the simulations, a proportion $p_s$ of them have strong first-stage ($\rho_g = 0.75$ and $\lambda_g = 281.25$). For the rest of the groups the instrument is not correlated with the endogenous variable ($\rho_g = 0$). The two graphs in the first column of Figure 1 plots 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 $\rho_g = 0.25$ and $\lambda_g = 31.25$. The two graphs in the second column plots the CDFs of DGP 2, where $p_s = p_w = 0.125$ and $p_s = p_w = 0.025$, respectively. DGP 3 represents a case where the weak and strong groups have no separation. Eighty percent of the groups in DGP 3 have irrelevant instruments. Among rest twenty percent of groups, half of them have first-stage effect $\rho_g \sim \mathcal{N}(0.1, 0.1^2)$ and the other half have $\rho_g \sim \mathcal{N}(0.75, 0.25^2)$. Note that all DGPs considered in this section have large proportions of zero groups, as is motivated by the data patterns of the two empirical examples in Section 4. If, in practice, the proportion of zero groups is small, the MSE gains of the proposed adaptive procedure could be negligible or even threatened by the risk of making type II error in finite samples. In such cases, the repeated split-sample fully-interacted estimator $\hat{\beta}_{ss,int}$ defined in (8) or the fully interacted LIML estimator that do not involve first-stage selection may be preferred.

We first present Monte Carlo evidence on the finite sample relevance of the asymptotic MSE approximations derived in Section 2 and show that the theoretical formula in Theorem 2 matches with simulation results. Figure 2 illustrate the case of DGP1 with $\rho_{u,v} = 0.3$ and $p_s = 0.25$. The total number of groups $G$ vary along the horizontal axis. The top graphs plot the empirical MSE of the full-interact ($\hat{\beta}_{int}$, FULLINT), select-and-interact ($\hat{\beta}_{sel,int}$ SELINT) and their associated repeated split sample estimators ($\hat{\beta}_{ss,int}$ FULLSSINT, $\hat{\beta}_{sssel,int}$ SELSSINT), with different cut-off $\delta$ value used in different columns. Results are based on 500 simulations. The bottom graphs plot the analytical formula of approximated asymptotic
MSE given in Section 2. We see that empirical MSEs in the top figures are closely matched by their theoretical counterparts. In addition, as predicted by the asymptotic approximation, the FULLINT and SELINT estimators have MSE curves that grow with $G$ at a quadratic rate.\(^5\) In contrast, the SELSSINT estimator has MSE that grows linearly with $G$. Because of the different growth rate with respect to $G$, when $G$ becomes large, the gain of using the repeated split sample estimation procedure becomes substantial.

**Figure 1:** Distribution of $\rho$ for three DGPs

\[^5\]In Figure 2 the MSE curves of FULLINT and SELINT look linear on the figure because $G^2/N \approx G/n_g$ and since $n_g$ is fixed in the simulation, the MSE curve grows linearly in $G$.\[^5\]
Figure 2: Empirical and Theoretical MSE Decomposition of Various Estimators
Figure 2 also report the MSE of the fully-interacted LIML estimator, which is known to have higher-order leading term of order $G/N$ in asymptotic approximation of MSE (Donald and Newey, 2001). For the three DGPs considered in the figure, LIML performs closely to the SELSSINT estimator. As we show below in additional simulations, LIML can perform better than the SELSSINT estimator for certain values of the parameters of the DGP but can also be dominated in other cases. Moreover, as in one of the applications in Section 4, LIML may produce an extremely large estimates for the parameter of interest, which highlights the drawback of LIML as an estimator that does not have finite sample moments.

In practice, it is useful to adopt an adaptive method to pick the value of $\delta$. Table 2 reports the empirical MSE of various estimators across 500 simulation experiments and the associated rejection proportions of the second-stage t-test under DGP 1. We find that the adaptive estimator proposed in Theorem 3 with $C = 0.1$ works very well under this well-separated DGP. Besides the FULLINT, the FULLSSINT, and the LIML estimators discussed above, Table 2 also reports simulation results for the pooled estimator $(\hat{\beta}_{\text{pool}}, \text{POOL})$, where first-stage heterogeneity is completely ignored, the infeasible repeated split-sample select-and-interact estimator (INFSSINT), which interacts the instrument only with groups that have non-zero first-stage correlation, and the adaptive estimator (ADPT) corresponding to the proposed repeated split-sample select-and-interact estimator with data-driven first-stage selection rule. For all three DGPs, the error terms are generated with either the normal or the $\chi^2_3$ distribution.

For normal errors, there is a clear gain from using the adaptive estimator, especially when the proportion of non-zero groups is relatively small (5 percent) such that it is of vital importance to exclude groups where the instrument does not provide identification in the first-stage. The adaptive estimator is almost as good as the corresponding infeasible oracle, or INFSSINT, which uses oracle information on the identity of the non-zero groups to construct a 2SLS estimator. LIML is a very competitive estimator overall, but its performance gets dominated by the adaptive estimator when the proportion of strong groups is small. Results are very similar when error distribution follows a $\chi^2$ distribution with three
degrees of freedom. While our analytical results are derived for the case of normal errors, the simulations in Table 2 indicate some robustness to alternative distributions.

Table 2: Rejection Proportion and MSE Performance of DGP 1

<table>
<thead>
<tr>
<th></th>
<th>POOL</th>
<th>FULLINT</th>
<th>FULLSSINT</th>
<th>INFSSINT</th>
<th>ADPT</th>
<th>LIML</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = 40</td>
<td>N×MSE</td>
<td>878.5594</td>
<td>41.4115</td>
<td>39.4421</td>
<td>35.2131</td>
<td>35.0878</td>
</tr>
<tr>
<td></td>
<td>Rej. Prop.</td>
<td>0.0440</td>
<td>0.0900</td>
<td>0.0580</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
<tr>
<td>G = 100</td>
<td>N×MSE</td>
<td>730.4278</td>
<td>51.7972</td>
<td>41.9885</td>
<td>36.2428</td>
<td>36.1591</td>
</tr>
<tr>
<td></td>
<td>Rej. Prop.</td>
<td>0.0400</td>
<td>0.1180</td>
<td>0.0500</td>
<td>0.0480</td>
<td>0.0500</td>
</tr>
<tr>
<td>G = 200</td>
<td>N×MSE</td>
<td>790.4691</td>
<td>75.7328</td>
<td>42.5080</td>
<td>37.5265</td>
<td>37.5073</td>
</tr>
<tr>
<td></td>
<td>Rej. Prop.</td>
<td>0.0520</td>
<td>0.2060</td>
<td>0.0500</td>
<td>0.0560</td>
<td>0.0560</td>
</tr>
<tr>
<td>G = 400</td>
<td>N×MSE</td>
<td>608.3507</td>
<td>105.9582</td>
<td>40.2268</td>
<td>35.3733</td>
<td>35.3776</td>
</tr>
<tr>
<td></td>
<td>Rej. Prop.</td>
<td>0.0300</td>
<td>0.3100</td>
<td>0.0440</td>
<td>0.0480</td>
<td>0.0480</td>
</tr>
</tbody>
</table>

|             | N×MSE | 30.2998  | 7.9778    | 7.7333   | 7.5790  | 7.5784  | 7.8078  |
|             | Rej. Prop. | 0.0560 | 0.0580    | 0.0500   | 0.0540  | 0.0540  | 0.0500  |
| G = 100     | N×MSE | 28.3546  | 8.4543    | 7.9890   | 7.8131  | 7.8109  | 7.9466  |
|             | Rej. Prop. | 0.0480 | 0.0680    | 0.0560   | 0.0520  | 0.0520  | 0.0640  |
|             | Rej. Prop. | 0.0600 | 0.0900    | 0.0540   | 0.0560  | 0.0560  | 0.0580  |
| G = 400     | N×MSE | 23.9340  | 10.2159   | 7.7901   | 7.7154  | 7.7107  | 7.6178  |
|             | Rej. Prop. | 0.0320 | 0.1040    | 0.0500   | 0.0540  | 0.0540  | 0.0480  |

|             | N×MSE | 411538.4388 | 221.4641  | 457.2013 | 240.9522 | 238.0250 | 331.8552 |
|             | Rej. Prop. | 0.0060 | 0.1040    | 0.0360   | 0.0420  | 0.0420  | 0.0860  |
| G = 100     | N×MSE | 27201.5543 | 314.0027  | 474.1837 | 229.5616 | 231.7304 | 287.8962 |
|             | Rej. Prop. | 0.0200 | 0.1820    | 0.0560   | 0.0540  | 0.0540  | 0.0860  |
| G = 200     | N×MSE | 5318.7396 | 492.6823  | 452.0095 | 233.8429 | 220.4919 | 312.9719 |
|             | Rej. Prop. | 0.0340 | 0.3340    | 0.0540   | 0.0520  | 0.0480  | 0.1020  |
| G = 400     | N×MSE | 4538.0426 | 872.5473  | 398.3639 | 212.9561 | 214.8910 | 294.0725 |
|             | Rej. Prop. | 0.0300 | 0.5900    | 0.0440   | 0.0320  | 0.0340  | 0.0920  |

Note: DGP1 under normal and $\chi^2_3$ errors. Scaled mean squared error and rejection probability are reported for different configurations of $G$, $p_s$. The adaptive estimator uses $C = 0.1$. 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

<table>
<thead>
<tr>
<th></th>
<th>POOL</th>
<th>FULLINT</th>
<th>FULLSSINT</th>
<th>INFSSINT</th>
<th>ADPT</th>
<th>LIML</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_s = p_w = 0.025, \rho_{u,v} = 0.3$ and normal errors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0380</td>
<td>0.1120</td>
<td>0.0420</td>
<td>0.0460</td>
<td>0.0440</td>
<td>0.0580</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>2925.0170</td>
<td>78.4621</td>
<td>78.6411</td>
<td>63.4108</td>
<td>62.8208</td>
<td>66.9665</td>
</tr>
<tr>
<td>$G = 100$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0360</td>
<td>0.2080</td>
<td>0.0420</td>
<td>0.0300</td>
<td>0.0480</td>
<td>0.0660</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>2193.4833</td>
<td>139.9809</td>
<td>104.7112</td>
<td>79.4951</td>
<td>77.0539</td>
<td>85.7289</td>
</tr>
<tr>
<td>$G = 200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0480</td>
<td>0.3100</td>
<td>0.0640</td>
<td>0.0540</td>
<td>0.0540</td>
<td>0.0740</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>1868.8223</td>
<td>175.9300</td>
<td>86.7155</td>
<td>67.6697</td>
<td>67.7630</td>
<td>73.8622</td>
</tr>
<tr>
<td>$G = 400$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0260</td>
<td>0.5120</td>
<td>0.0400</td>
<td>0.0420</td>
<td>0.0400</td>
<td>0.0500</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>1395.2831</td>
<td>268.4904</td>
<td>77.0749</td>
<td>60.9046</td>
<td>60.6703</td>
<td>68.6337</td>
</tr>
</tbody>
</table>

$p_s = p_w = 0.125, \rho_{u,v} = 0.3$ and normal errors

<table>
<thead>
<tr>
<th></th>
<th>POOL</th>
<th>FULLINT</th>
<th>FULLSSINT</th>
<th>INFSSINT</th>
<th>ADPT</th>
<th>LIML</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G = 40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0540</td>
<td>0.0600</td>
<td>0.0500</td>
<td>0.0480</td>
<td>0.0480</td>
<td>0.0700</td>
</tr>
<tr>
<td>$G = 100$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0440</td>
<td>0.1000</td>
<td>0.0620</td>
<td>0.0600</td>
<td>0.0680</td>
<td>0.0680</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>66.4858</td>
<td>17.2662</td>
<td>15.9882</td>
<td>15.2821</td>
<td>15.2456</td>
<td>15.6411</td>
</tr>
<tr>
<td>$G = 200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0580</td>
<td>0.1200</td>
<td>0.0540</td>
<td>0.0600</td>
<td>0.0600</td>
<td>0.0620</td>
</tr>
<tr>
<td>$G = 400$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0320</td>
<td>0.1440</td>
<td>0.0520</td>
<td>0.0460</td>
<td>0.0520</td>
<td>0.0520</td>
</tr>
</tbody>
</table>

$p_s = p_w = 0.025, \rho_{u,v} = 0.3$ and $\chi^2_3$ errors

<table>
<thead>
<tr>
<th></th>
<th>POOL</th>
<th>FULLINT</th>
<th>FULLSSINT</th>
<th>INFSSINT</th>
<th>ADPT</th>
<th>LIML</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G = 40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0040</td>
<td>0.1420</td>
<td>0.0240</td>
<td>0.0300</td>
<td>0.0320</td>
<td>0.1360</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>1715411.9228</td>
<td>379.2813</td>
<td>1938.4813</td>
<td>519.8839</td>
<td>535.2309</td>
<td>1806.3409</td>
</tr>
<tr>
<td>$G = 100$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0120</td>
<td>0.2780</td>
<td>0.0360</td>
<td>0.0460</td>
<td>0.0400</td>
<td>0.1260</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>459815.1212</td>
<td>676.1290</td>
<td>2186.4904</td>
<td>547.4009</td>
<td>526.7137</td>
<td>826.4837</td>
</tr>
<tr>
<td>$G = 200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0240</td>
<td>0.4400</td>
<td>0.0400</td>
<td>0.0480</td>
<td>0.0420</td>
<td>0.1560</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>16785.7984</td>
<td>949.1267</td>
<td>1146.9917</td>
<td>430.9470</td>
<td>445.8067</td>
<td>723.5287</td>
</tr>
<tr>
<td>$G = 400$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0260</td>
<td>0.7540</td>
<td>0.0460</td>
<td>0.0300</td>
<td>0.0380</td>
<td>0.1160</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>12458.9357</td>
<td>1732.1899</td>
<td>980.8136</td>
<td>384.3596</td>
<td>402.7681</td>
<td>654.9738</td>
</tr>
</tbody>
</table>

$p_s = p_w = 0.125, \rho_{u,v} = 0.3$ and $\chi^2_3$ errors

<table>
<thead>
<tr>
<th></th>
<th>POOL</th>
<th>FULLINT</th>
<th>FULLSSINT</th>
<th>INFSSINT</th>
<th>ADPT</th>
<th>LIML</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G = 40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0280</td>
<td>0.0720</td>
<td>0.0520</td>
<td>0.0540</td>
<td>0.0560</td>
<td>0.0640</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>368.2421</td>
<td>80.3004</td>
<td>100.1004</td>
<td>85.7815</td>
<td>89.1575</td>
<td>89.6680</td>
</tr>
<tr>
<td>$G = 100$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0360</td>
<td>0.1160</td>
<td>0.0520</td>
<td>0.0400</td>
<td>0.0380</td>
<td>0.0560</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>393.8971</td>
<td>103.9424</td>
<td>108.1042</td>
<td>82.5056</td>
<td>82.9206</td>
<td>84.7621</td>
</tr>
<tr>
<td>$G = 200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0500</td>
<td>0.1740</td>
<td>0.0460</td>
<td>0.0360</td>
<td>0.0520</td>
<td>0.0540</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>411.4270</td>
<td>132.1043</td>
<td>100.0079</td>
<td>79.3227</td>
<td>78.4769</td>
<td>84.7178</td>
</tr>
<tr>
<td>$G = 400$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rej. Prop.</td>
<td>0.0360</td>
<td>0.3380</td>
<td>0.0360</td>
<td>0.0540</td>
<td>0.0580</td>
<td>0.0580</td>
</tr>
<tr>
<td>$N \times MSE$</td>
<td>373.1543</td>
<td>214.7361</td>
<td>94.7822</td>
<td>81.5538</td>
<td>88.8683</td>
<td>81.6769</td>
</tr>
</tbody>
</table>

Note DGP2 under normal and $\chi^2_3$ errors. Scaled mean squared error and rejection probability are reported for different configurations of $G$, $p_s$, and $p_w$. The adaptive estimator uses $C = 0.1$. The group sample size is fixed at $n_g = 500$. Results are based on 500 simulation repetitions.

Table 3 reports results for DGP 2, again with both normal and $\chi^2_3$ error distributions. Under this DGP, groups with weak first-stage identification are added to the specification. Recall that the INFSSINT estimator uses oracle information and interacts the instrument with groups that have a non-zero first-stage correlation. The estimator is no longer optimal under DGP 2 because excluding some weak groups can help bring down MSE of the...
second-stage estimation through bias reduction. Therefore, we see in Table 3 that the ADPT estimator can sometimes have the best empirical MSE among all reported estimation methods. In other cases, the ADPT estimator performs very closely to the INFSSINT estimator. Another interesting discovery in Table 3 is that the LIML estimator can have unfavorable small sample performance in both empirical MSE and size control when the proportion of groups with first-stage unidentification is large and when the error term distribution is far from normal. The performance of LIML improves when the proportion of groups with first-stage unidentification decreases.

Table 4: Rejection Proportion and MSE Performance for DGP 3

<table>
<thead>
<tr>
<th></th>
<th>POOL</th>
<th>FULLINT</th>
<th>FULLSSINT</th>
<th>INFSSINT</th>
<th>ADPT</th>
<th>LIML</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = 40</td>
<td>N×MSE</td>
<td>302.1607</td>
<td>35.0643</td>
<td>34.6589</td>
<td>32.3089</td>
<td>32.3214</td>
</tr>
<tr>
<td></td>
<td>Rej. Prop.</td>
<td>0.0520</td>
<td>0.1020</td>
<td>0.0640</td>
<td>0.0680</td>
<td>0.0660</td>
</tr>
<tr>
<td></td>
<td>Rej. Prop.</td>
<td>0.0440</td>
<td>0.0800</td>
<td>0.0480</td>
<td>0.0420</td>
<td>0.0440</td>
</tr>
<tr>
<td></td>
<td>Rej. Prop.</td>
<td>0.0560</td>
<td>0.1200</td>
<td>0.0560</td>
<td>0.0600</td>
<td>0.0640</td>
</tr>
<tr>
<td></td>
<td>Rej. Prop.</td>
<td>0.0300</td>
<td>0.1700</td>
<td>0.0460</td>
<td>0.0500</td>
<td>0.0460</td>
</tr>
</tbody>
</table>

DGP3 with $\rho_{u,v} = 0.3$ and normal errors

| G = 40           | N×MSE | 2354.2316 | 237.1810  | 315.0052  | 195.8793 | 206.1994 | 240.4183 |
|                  | Rej. Prop. | 0.0220 | 0.1760 | 0.0420 | 0.0440 | 0.0420 | 0.0860 |
| G = 100          | N×MSE | 1187.5111 | 244.3812  | 187.9446  | 129.5377 | 131.7042 | 137.1416 |
|                  | Rej. Prop. | 0.0360 | 0.2660 | 0.0680 | 0.0580 | 0.0580 | 0.0780 |
| G = 200          | N×MSE | 925.6875 | 280.9925  | 132.5488  | 93.9221 | 93.2734 | 102.7922 |
|                  | Rej. Prop. | 0.0500 | 0.3940 | 0.0440 | 0.0400 | 0.0380 | 0.0660 |
| G = 400          | N×MSE | 828.6069 | 534.7148  | 130.5877  | 103.2260 | 101.4422 | 105.6785 |
|                  | Rej. Prop. | 0.0360 | 0.6940 | 0.0440 | 0.0500 | 0.0580 | 0.0620 |

Note: DGP3 under normal and $\chi^2_3$ errors. Scaled mean squared error and rejection probability are reported for different configurations of $G$. The adaptive estimator uses $C = 0.1$. The group sample size is fixed at $n_g = 500$. Results are based on 500 simulation repetitions.

Table 4 reports simulation results for DGP3, where groups with weak and strong first-stage effects are not well-separated. For this DGP, we redefine the INFSSINT estimator as the infeasible estimator that chooses $\delta$ to minimize the theoretical MSE of the repeated split-sample select-and-interact estimator stated in Thereom 2 using oracle information of $\rho_g$. The ADPT estimator again has performance that is very close to the infeasible estimator. Also as in Table 2, the ADPT estimator sees substantial gains in MSE compared to its fully...
interacted counterpart when the proportion of non-zero group is relatively small, and the performance seems to be robust to whether the error terms are jointly normally distributed or have distributions with fat tails.

4. Empirical Examples

4.1. Return to Compulsory Schooling

The return to compulsory schooling literature studies how an extra year of (compulsory) schooling affects individual well-being, such as earnings and health outcomes, later in life. Researchers have often exploited variation in compulsory schooling laws across states and overtime in the U.S. (see, Lleras-Muney, 2005, Oreopoulos, Page, and Stevens, 2006, and Stephens and Yang, 2014, among others) and other countries (Oreopoulos, 2006) to instrument endogenous schooling choice that may correlate with omitted variables, such as early cognitive ability and family background. The argument for identification is that any law change in minimum school leaving age affects individual education attainment, but not individual well-being later in life, other than through the education channel.

This section revisits the public-use U.S. Census dataset compiled by Oreopoulos (2006) while considering the possibility of heterogeneous effect in the compulsory law changes in the first stage. The dataset records cell means of variables for individuals collapsed by birth state, birth cohort, census year, gender, and race. Let subscript $s$ denote birth state, $t$ denote birth cohort, $c$ denote census year, and $d$ denote demographic characteristics gender and race. Formally, we consider the following specification.

$$\text{Logwage}_{stcd} = \beta \text{Educ}_{stcd} + X_{stcd}\theta + u_{stcd}$$

$$\text{Educ}_{stcd} = \sum_{g} \rho_{g} I(S_{stcd} = g)\text{CL}_{st} + X_{stcd}\gamma + v_{stcd}$$

where $\text{Logwage}_{stcg}$ and $\text{Educ}_{stcg}$ are the log wage and years of schooling, while $\text{CL}_{st}$ is the compulsory schooling year a birth cohort faces at age 14 if born in year $t$ and state $s$. The regressor $X_{stcd}$ includes fixed effects for survey year, birthstate, gender, race, a quartic in age, and state-specific birth-year time trend (specification in Table 4, Column 3 of Oreopoulos,
The first stage correlation between the instrument and the endogenous regressor is allowed to be heterogeneous across groups. We use three different definitions for groups, 1) by state only, 2) by state and gender, and 3) by state, gender, and race (dummy variable for whether an individual is Black), and 4) by state, gender, and race (dummy variable for whether an individual is Black). States are natural groups in our context because the minimum school leaving age in the U.S. is determined by state-specific laws and regulations. Gender is used because the literature has found that males are most responsive to changes in the minimum school leaving age and often single them out for subsample analysis. Besides males, Oreopoulos (2006) also selects Black males for subsample analysis, while Stephens and Yang (2014) selects White males.

Figure 3 illustrates the first-stage strength when identification relies on variations in minimum school leaving age. Groups are defined solely by birthstate. The left graph plots the $p$-values of groupwise first-stage upper one-sided $t$ tests for instrument strength. We see in all graphs that there is a fraction of groups with $p$-values that are essentially zero and a fraction of groups with $p$-values following a close-to-uniform distribution, suggesting a mixture of groups with strong and irrelevant IV. The right graph plots the top ten groupwise $\hat{\mu}_g$ estimates against their corresponding $p$-values reported in the left graph. We see that although there are a dozen states with close to zero $p$-values, the selection metric of our proposed adaptive procedure, $\hat{\mu}_g$, flattens out after a handful of states. Considering the trade-off between higher-order bias and efficiency loss, our adaptive procedure selects the first five stages for second-stage estimation, as is also reported in Panel A of Table 5.

---

6Oreopoulos (2006) also considered a model specification with additional state and birth cohort level control variables such as fractions living in urban areas, black, in the labor force, in the manufacturing sector, female, and average age calculated based on when a birth cohort was age 14. Oreopoulos (2006) found that the 2SLS estimator is insignificant for the full sample but statistically significant at the 5 percent level for the Black males. However, a closer look at group-wise first stage testing results similar to those displayed in Figures 3 and 4 suggests lack of identification due to near multicollinearity.
Figure 3: Return to Compulsory Schooling: First-stage Signals by State

Note: Dataset is from Oreopoulos (2006). Groups with no variation in the instrument are dropped. The endogenous regressor is years of schooling. The instrument is the compulsory schooling year a birth cohort faces at age 14. Other exogenous controls include fixed effects for survey year, state, gender, race, a quartic in age, and state-specific birth-year time trend. The left graph plots p-values of groupwise first-stage t-tests. The second graph plots the top ten groupwise $\hat{\mu}_g$ estimates against their corresponding first-stage p-values.

Figure 4 repeats the groupwise t-tests for instrument strength using alternative group definitions as discussed above. We see similar evidence of mixing between groups with strong and irrelevant IV.

Figure 4: Return to Compulsory Schooling: p-values of Groupwise First-Stage t-Tests

Note: Dataset is from Oreopoulos (2006). Groups with no variation in the instrument are dropped. When groups are defined by state, gender, and race, 18 groups of male or female Blacks in small states are also dropped due to insufficient observations in earlier Census surveys. The endogenous regressor is years of schooling. The instrument is the compulsory schooling year a birth cohort faces at age 14. Other exogenous controls include fixed effects for survey year, state, gender, race, a quartic in age, and state-specific birth-year time trend.

Table 5 reports estimation results. In the first panel, the data is grouped by birthstate. After dropping states that do not have variation in minimum school leaving age, there are 33
groups left for analysis. Pooled 2SLS in Column 2 estimates an effect of 17.3 percent increase in wage for an additional year of schooling, which is comparable to the results reported in Table 4 Column 3 of Oreopoulos (2006). Fully interacted 2SLS in Column 3 estimates an effect of 8.6 percent. However, noticing that the OLS estimate (Column 1) is 9 percent, it is likely that the smaller estimate originates from the “many IV bias” which is toward the direction of OLS. The proposed adaptive estimator selects California, Delaware, Louisiana, New Mexico, and Ohio out of thirty-three states and gives an estimated effect of 10.3 percent. The estimator also reports a 24 percent reduction in standard errors compared to the pooled 2SLS estimator. Table 5 also reports estimates from the fully interacted LIML (Column 4) and fully interacted repeated split-sample estimator, defined in (8) (Column 5, abbreviated as RSS). The LIML estimator is out of reasonable range, likely due to the fact that LIML has no finite sample moments regardless of the number of instruments and could potentially have extreme estimates. The fully interacted repeated split-sample estimator estimates an effect of 9.6 percent increase in wage for an additional year of schooling with a smaller standard error than the adaptive estimator. This is expected as the adaptive estimator is designed to improve the MSE by taking some (higher-order) efficiency loss in return for bias reduction.

Table 5: Return to Compulsory Schooling: Estimation Results

<table>
<thead>
<tr>
<th></th>
<th>Pool</th>
<th>Fully Interacted</th>
<th>Proposed ADPT 2SLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>2SLS</td>
<td>2SLS</td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Panel A: groups defined by state</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.090</td>
<td>0.173</td>
<td>0.086</td>
<td>0.562</td>
</tr>
<tr>
<td>(0.002)</td>
<td>(0.068)</td>
<td>(0.021)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>Panel B: groups defined by state and gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.090</td>
<td>0.173</td>
<td>0.103</td>
<td>0.589</td>
</tr>
<tr>
<td>(0.002)</td>
<td>(0.068)</td>
<td>(0.192)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>Panel C: groups defined by state, gender, and race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.094</td>
<td>0.159</td>
<td>0.116</td>
<td>0.580</td>
</tr>
<tr>
<td>(0.002)</td>
<td>(0.061)</td>
<td>(0.017)</td>
<td>(0.003)</td>
</tr>
</tbody>
</table>

Note: Dataset is from Oreopoulos (2006). Groups with no variation in the instrument are dropped. For the third grouping rule, 18 groups of male or female Blacks in small states are also dropped due to insufficient observations in earlier surveys. The outcome variable is log wage. The endogenous regressor is years of schooling. The instrument is the compulsory schooling year a birth cohort faces at age 14. Other exogenous controls include fixed effects for birth-year, state, survey year, gender, race, a quartic in age, and state-specific birth-year time. The adaptive estimator uses $C = 0.1$. The method selects females from DE and males from CA, IA, LA, ME, MI, NM in Panel B. It selects white males from CA, DE, NM, UT, black males from GA, IA, LA, MI, NC, SC, and black females from CA, DE, LA, OH, OK in Panel C.
In the second panel, the data is grouped by both birthstate and gender, and there is a total of 66 groups. The proposed adaptive estimator selects 7 out of the 66 groups and estimates an effect of 15.1 percent wage increase. The fully interacted 2SLS estimator as well as its repeated split-sample version are both insignificant in this panel, while the fully interacted LIML estimator again reports a large effect estimate of over 50 percent wage increase. In addition, it is worth noting that six out of the seven selected groups are men, which is in line with the previous literature which documented that men responded more to law changes in minimum school leaving age.

In the third panel, the data is grouped by birthstate, gender, and race, and there are a total of 114 groups. 18 groups of male or female Blacks in small states are dropped due to insufficient observations in earlier surveys. The fifteen selected groups consist of four groups of White males, six groups of Black males, and five groups of Black females. No White female groups are selected. In this panel, the proposed adaptive estimator again consistently reports an estimated effect of a large 19.6 percent wage increase.

4.2. Voter Turnout

Charles and Stephens (2013) use county-level data to study the effect of local market activity 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 presidents. The identification strategy is first to difference out county-level fixed effects and then account for potentially endogenous changes in local market activities using exogenous shocks to oil/natural gas (oil, thereafter) and coal supply. The identification strategy follows from Black (2002) who utilize coal shocks to study the impact of local economic conditions on participation in programs of disability payments, and Acemoglu and Notowidigdo (2013) who utilize oil shocks to study the effect of local income on health spending. Recently, Charles, Li, and Stephens (2018) uses oil shocks to study the effect of local labor market conditions on disability take-up in federal programs.

The instruments used in the above-mentioned literature are typically a measure of exogenous energy supply shocks, including changes in oil or coal price or changes in national
employment in oil or coal production industries, interacted with a measure of the importance of coal/oil industry in a county prior to the period of study. As is expected, the identification power of the first-stage instrument varies across states, and the aforementioned papers often restrict the data sample to a pre-determined list of oil or coal states. For example, Charles and Stephens (2013) define coal states to be Kentucky, Ohio, Pennsylvania, and West Virginia following Black (2002), and define 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) also add Louisiana to the list of oil states. Acemoglu and Notowidigdo (2013) use the sample of southern states.

In this section, we revisit the empirical questions studied in Charles and Stephens (2013). We adopt a similar identification strategy as theirs, except that we do not pre-specify the coal and oil states nor the counties with larger coal or oil presence. Instead, we use the proposed adaptive estimation strategy to identify states with higher first stage power with the goal of minimizing asymptotic MSE of the second stage IV estimator. Since we do not specify exclusive lists of coal and oil states –in fact, our analysis shows that some states are affected by both supply shocks– we will use coal and oil shocks as different IV specifications. We will also explore different definitions of coal and oil shocks.

Let subscript \( c \) denote county, \( s \) denote state, and \( t \) denote year (when an election takes place). Specifically, the model is specified as follows,

\[
\Delta Vote_{cst} = \beta \Delta Earnings_{cst} + X_{cst} \theta + u_{cst}
\]

\[
\Delta Earnings_{cst} = \sum_{s=1}^{S} \rho_s \Delta EnergySupply_t \times EnergyShare_{cs} + X_{cst} \gamma + v_{cst},
\]

where \( \Delta Vote_{cst} \) is the change in voter turnout between two elections, \( \Delta Earnings_{cst} \) is the change in log per capital earning, the instrument \( \Delta EnergySupply_t \times EnergyShare_{cs} \) is the change in oil/gas or coal supply shock measured by national employment level interacted with county-level employment share in such an industry, and \( X_{cst} \) 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. Since the data spans from 1969 to 1990, ideally the \textit{EnergyShare}_{cs} variable needs to be measured before 1969. Unfortunately, the 1967 CBP only offers a crude employment measurement for the mining industry, and therefore Charles and Stephens (2013) use the 1974 CBP employment share, which distinguishes oil production form coal production, as their benchmark measurement for local energy production.

The following four graphs in Figure 5 reports the \( p \)-values of group-wise first-stage \( t \) tests using the model specification discussed above and four different measurements of the instrument. We see that in all four graphs there is strong evidence of mixing in the strength of first-stage signal as is expected because states with little presence of the oil or coal industry might see little variation in the instrument variable and hence very noisy signal in the first stage. As is discussed in the theoretical section, both the variation in the instrument and the first stage coefficient \( \rho_{y} \) determines the contribution of a specific group to the precise estimation of the second stage \( \beta \) coefficient. This provides ample motivation to apply our proposed methodology of selecting strong first stage signals with the target of the asymptotic MSE of the split-sample select-and-interact estimator.

Table 6 reports regression results for the effect of local earnings on voter turnout in gubernatorial elections. The first column reports the pooled 2SLS estimates using data samples from oil or coal states specified in Charles and Stephens (2013). The second to fourth columns report various full-sample IV estimates using the whole set of instruments interacted with state dummies. The fifth column reports the proposed adaptive 2SLS estimates and the states by the adaptive procedure. First we notice that all the adaptive and fully interacted IV estimators reported in this table are negative and statistically significant and are in line with the result in Charles and Stephens (2013). That is, better local economic conditions lower turnout in gubernatorial elections. However, the pooled 2SLS estimates using coal states alone are not statistically different from zero. This is likely because there are only four pre-determined coal states: Kentucky, Ohio, Pennsylvania, and West Virginia. The list
Figure 5: Gubernatorial 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 employment per adult since last election of governor. The instrument is change in national employment in oil or coal interacted with share of oil or coal production in local employment. 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.
of coal states also does not turn out to overlap much with states selected by the adaptive procedures. On the other hand, the list of oil states in Charles and Stephens (2013) mostly includes all states selected by the adaptive procedure with either 1967 or 1974 CBP measures of local employment share.

Table 6: Gubernatorial Voter Turnout: Estimation Results

<table>
<thead>
<tr>
<th></th>
<th>Oil/Coal States Only</th>
<th>Fully Sample, Fully Interacted</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pooled 2SLS</td>
<td>2SLS</td>
<td>LIML</td>
</tr>
<tr>
<td>Panel A: oil supply shock interacted with 1974 CBP local employment share</td>
<td>-0.085 (0.025)</td>
<td>-0.063 (0.015)</td>
<td>-0.074 (0.017)</td>
</tr>
<tr>
<td>Panel B: coal supply shock interacted with 1974 CBP local employment share</td>
<td>0.011 (0.040)</td>
<td>-0.030 (0.021)</td>
<td>-0.036 (0.023)</td>
</tr>
<tr>
<td>Panel C: oil supply shock interacted with 1967 CBP local employment share</td>
<td>-0.050 (0.025)</td>
<td>-0.044 (0.015)</td>
<td>-0.050 (0.016)</td>
</tr>
<tr>
<td>Panel D: coal supply shock interacted with 1967 CBP local employment share</td>
<td>0.034 (0.051)</td>
<td>-0.055 (0.014)</td>
<td>-0.062 (0.015)</td>
</tr>
</tbody>
</table>

Note: Dataset is from Charles and Stephens (2013). The outcome variable is the change in voter turnout since the last election of governor, where voter turnout is calculated by dividing the total number of votes using the number of individuals 20 years old and above residing in the county. The endogenous regressor is the change in log county-level employment per adult. The instrument is the change in national employment in oil or coal interacted with the share of oil or coal production in local employment. 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. The adaptive estimator is calculated with constant $C = 0.1$.

Next we repeat the exercise for senate elections and report estimation results in Figure 6 and Table 7. We find similar estimation results compared to the Gubernatorial elections when oil supply shock is used, with similar sets of states selected by the adaptive procedure. However, when coal supply shocks are used as instruments, a very different set of states are chosen by the adaptive procedure, and the estimates are mixed both in sign and in statistical significance.
Note: Dataset is from Charles and Stephens (2013). The endogenous regressor is change in log county-level employment per adult since last election of senator. The instrument is the change in national employment in oil or coal interacted with share of oil or coal production in local employment. 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. The adaptive estimator uses $C = 0.1$. 

34
Table 7: Senatoral Voter Turnout: Estimation Results

<table>
<thead>
<tr>
<th>Oil/Coal States Only</th>
<th>Fully Sample, Fully Interacted</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled 2SLS</td>
<td>2SLS</td>
<td>LIML</td>
</tr>
<tr>
<td>Panel A: oil supply shock interacted with 1974 CBP local employment share</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.077</td>
<td>-0.067</td>
<td>-0.098</td>
</tr>
<tr>
<td>(0.029)</td>
<td>(0.017)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Panel B: coal supply shock interacted with 1974 CBP local employment share</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.000</td>
<td>0.025</td>
<td>0.037</td>
</tr>
<tr>
<td>(0.026)</td>
<td>(0.017)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Panel C: oil supply shock interacted with 1967 CBP local employment share</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.047</td>
<td>-0.038</td>
<td>-0.055</td>
</tr>
<tr>
<td>(0.023)</td>
<td>(0.014)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Panel D: coal supply shock interacted with 1967 CBP local employment share</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.035</td>
<td>-0.022</td>
<td>-0.029</td>
</tr>
<tr>
<td>(0.032)</td>
<td>(0.012)</td>
<td>(0.014)</td>
</tr>
</tbody>
</table>

Note: Dataset is from Charles and Stephens (2013). The outcome variable is the change in voter turnout since the last election of senators, where voter turnout is calculated by dividing the total number of votes using the number of individuals 20 years old and above residing in the county. The endogenous regressor is the change in log county-level employment per adult. The instrument is the change in national employment in oil or coal interacted with the share of oil or coal production in local employment. 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. The adaptive estimator is calculated with constant $C = 0.1$. 

35
Appendix A: Proofs of Lemmas, Theorems, and Corollaries

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

Proof. The threshold $\delta^*$ corresponding to the Bonferroni-type multiple testing controlling familywise error rate $\alpha$ 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 $1 - \Phi(\delta^*/\sigma_v) > 0.5$, we have that $\frac{\delta^*}{\sigma_v} \leq \phi(\delta^*)/\delta^*$. For any $\delta^* > 1$, we further have that $\frac{\delta^*}{\sigma_v} \leq \phi(\delta^*) \leq \exp(-\delta^*^2/2)$. Therefore, $\delta^* = O(\sqrt{\log G})$, which implies that $\delta^* = o((N/G)^{1/2})$ 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 $\delta$ corresponding to those procedures will not be larger than the Bonferroni method.

Proof of Some Useful Results

We replace $c_{g, \alpha_{FS}}$ by $c_g$ and $\alpha_{FS}$ by $\alpha$ in all following proofs for notational convenience. 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 $C < \infty$ such that $A_g \leq C \cdot n_g$ for all $g = 1, \ldots, G$. Also, let $\tilde{X}_g = [Z_g \ X_g]$ be an $n_g \times (d + 1)$ matrix. For all $g = 1, \ldots, G$, let $H_{g,1} = \sqrt{W_g'M_gW_g/n_g}$ and $H_{g,2} = \sqrt{Z'_gZ_g/n_g}$.

Let $k = \lim_{G,N \to \infty} \frac{1}{N} \sum_{g=1}^{G} Z'_gZ_g \to k$, $k' = \lim_{G,N \to \infty} \frac{1}{N} \sum_{g=1}^{G} \rho_g Z'_gZ_g$, and $k'' = \lim_{G,N \to \infty} \frac{1}{N} \sum_{g=1}^{G} \rho_g^2 Z'_gZ_g$. By Assumption 1, we also know that $\frac{1}{N} \sum_{g=1}^{G} \rho_g Z'_gZ_g \to k'$ and $\frac{1}{N} \sum_{g=1}^{G} \rho_g^2 Z'_gZ_g \to k''$ as the corresponding averages across the other two groups are $\alpha_p(1)$.

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 $H_{g,1}^2 = W_g M_{X_g} W_g/n_g = v'_g M_{\tilde{X}} v_g/n_g = v'_g v_g/n_g - v'_g P_{\tilde{X}_g} v_g/n_g$, where $E[v'_g P_{\tilde{X}_g} v_g|\tilde{X}_g] = tr(P_{\tilde{X}_g})E[v'_g v_g|\tilde{X}_g] = \sigma^2_v(d+1)$. We know that

$$P(|H_{g,1}^2 - \sigma_v^2| > \sigma_v^2/2) \leq P(|v'_g v_g/n_g - \sigma_v^2| + v'_g P_{\tilde{X}_g} v_g/n_g > \sigma_v^2/2)$$

$$= P(|v'_g v_g/n_g - \sigma_v^2| > \sigma_v^2/2 - v'_g P_{\tilde{X}_g} v_g/n_g, v'_g P_{\tilde{X}_g} v_g/n_g < \sigma_v^2/2) + P(|v'_g v_g/n_g - \sigma_v^2| > \sigma_v^2/2 - v'_g P_{\tilde{X}_g} v_g/n_g, v'_g P_{\tilde{X}_g} v_g/n_g > \sigma_v^2/2)$$

$$\leq P(|v'_g v_g/n_g - \sigma_v^2| > \sigma_v^2/4) + P((v'_g P_{\tilde{X}_g} v_g/n_g - \sigma_v^2(d+1)/n_g) > \sigma_v^2/2 - \sigma_v^2(d+1)/n_g)$$

$$\leq P(|v'_g v_g/n_g - \sigma_v^2| > \sigma_v^2/4) + P(v'_g P_{\tilde{X}_g} v_g/n_g - \sigma_v^2(d+1)/n_g > \sigma_v^2/4)$$

$$\leq E[(\sum_{i=1}^{n_g} v_{ig}^2/n_g - \sigma_v^2)^4]/(\sigma_v^4)^4 + E[(v'_g P_{\tilde{X}_g} v_g)^2]/n_g^2/(\sigma_v^2/4)^2),$$

where the second to the last inequality holds for $n_g \geq 4(d+1)$, and the last line is from Chebyshev’s inequality. Further, we have that

$$E[(\sum_{i=1}^{n_g} v_{ig}^2/n_g - \sigma_v^2)^4] = V[\sum_{i=1}^{n_g} v_{ig}^2/n_g]^2 + V[\sum_{i=1}^{n_g} v_{ig}^2/n_g - \sigma_v^2]^2$$

$$= V[v_{ig}^2/n_g]^2 + V[\sum_{i=1}^{n_g} (v_{ig}^2/n_g - \sigma_v^2)] = V[v_{ig}^2/n_g]^2 + V[(v_{ig}^2 - \sigma_v^2)^2]/n_g^3$$

$$\leq 1/n_g^2,$$

due to moment conditions of the error term $v_{ig}$ in Assumption 1. To bound $E[(v'_g P_{\tilde{X}_g} v_g)^2]$, let $P_{ij}$ denote the $(i, j)$-th element of $P_{\tilde{X}_g}$ and notice that $\sum_{i=1}^{n_g} P_{ii} = tr(P_{\tilde{X}_g}) = d + 1$, $0 \leq P_{ii} \leq 1$, and 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 $0 \leq \sum_{i \neq j} P_{ij} P_{ij} \leq \sum_{i, j} P_{ij}^2 = tr(P_{\tilde{X}_g} P_{\tilde{X}_g}) = d + 1$. Further, one could show that $P_{ii} \geq 1/n_g$ and further tightening some of the bounds above if necessary. Then we know

$$E[(v'_g P_{\tilde{X}_g} v_g)^2|X_g] = \sum_{i,j,k,l} E[v_{ig} P_{ij} v_{jk} P_{kl} v_{lg}|X_g]$$

$$= \sum_i P_{ii} E(v_{ig}^4|X_g) + \sum_{i \neq j} P_{ii} P_{jj} E(v_{ig}^2 v_{ij}^2|X_g)$$

$$+ \sum_{i \neq j} P_{ij} E(v_{ig}^2 v_{ij}^2|X_g) + \sum_{i \neq j} P_{ij} E(v_{ij}^2 v_{ij}^2|X_g)$$

37
is also bounded with a universal constant across \( g = 1, ..., G \) due to the moment conditions of \( \nu_{ig} \).

Notice that \( H^2_{g,2} = \tilde{Z}_g' M_{X_g} \tilde{Z}_g / n_g \). So the second inequality could be proven following the same arguments as above with details omitted.

**Lemma A3.** Under Assumption 1 and provided that \( \delta \leq C_\delta (\frac{N}{G})^{1/2} \) with \( C_\delta < \rho \sqrt{kc} \), then for all \( g \in G_{+,s} \)

\[
P(\hat{\mu}_g \leq \delta) \lesssim \frac{1}{n_g^2}.
\]

**Proof.** Since \( C_\delta < \rho \sqrt{kc} \), then there exists a small positive constant \( \eta \in (0, 1) \) such that \( C_\delta \leq \rho \sqrt{kc} (1 - \eta) \). Given that \( \eta \), we know

\[
P(\hat{\mu}_g \leq \delta) = P \left( \frac{Z_g' \nu_g}{\sqrt{Z_g' Z_g}} \leq \delta - \mu_g \right)
\]

\[
= P \left( \frac{Z_g' \nu_g}{\sqrt{Z_g' Z_g}} \leq \delta - \mu_g; \delta \geq \sqrt{1 - \eta} \cdot \mu_g \right) + P \left( \frac{Z_g' \nu_g}{\sqrt{Z_g' Z_g}} \leq \delta - \mu_g; \delta < \sqrt{1 - \eta} \cdot \mu_g \right)
\]

\[
\leq P \left( \delta \geq \sqrt{1 - \eta} \cdot \mu_g \right) + P \left( \frac{Z_g' \nu_g}{\sqrt{Z_g' Z_g}} \leq (\sqrt{1 - \eta} - 1) \mu_g \right).
\]

For all \( g \in G_{+,s} \), based on similar argument as in Lemma A2, we know that

\[
P \left( \frac{Z_g' \nu_g}{\sqrt{Z_g' Z_g}} \leq (\sqrt{1 - \eta} - 1) \mu_g \right) \lesssim \frac{1}{n_g^2},
\]

and that

\[
P \left( \delta \geq \sqrt{1 - \eta} \cdot \mu_g \right) = P \left( \sqrt{Z_g' Z_g} \leq \delta / \rho_g / \sqrt{1 - \eta} \right)
\]

\[
\leq P \left( \sqrt{Z_g' Z_g} \leq \rho / \sqrt{kc} (1 - \eta) (\frac{N}{G})^{1/2} / \rho_g / \sqrt{1 - \eta} \right)
\]

\[
\leq P \left( \sqrt{Z_g' Z_g} < \sqrt{k_g n_g (1 - \eta)} \right) \lesssim \frac{1}{n_g^2},
\]

where the last line follows from the result that with similar arguments as in Lemma A2, one can show that \( P \left( |H^2_{g,2} - k_g| > \eta \cdot k_g \right) \lesssim \frac{1}{n_g^2} \) for any fixed positive \( \eta \). The lemma is then proven.

\( \square \)
Lemma A4. Under Assumption 1, for $\delta \leq C_3(NG)^{1/2}$ with $C_3 < \rho \sqrt{k}$ and any non-negative integer $k$, we have that as $G, N \to \infty$

1. $|((\delta - \mu_g)/\sigma_v|^k \phi((\delta - \mu_g)/\sigma_v) \leq (k/2)!/(\sqrt{2\pi}\sigma_v) < \infty$ for all $g = 1, \ldots, G$;

2. $E[(\delta - \mu_g)/\sigma_v|^k \phi((\delta - \mu_g)/\sigma_v) \leq \frac{1}{\eta^n}$ and $E[\Phi((\delta - \mu_g)/\sigma_v)] \leq \frac{1}{\eta^n}$ for all $g \in G_{+, s}$;

3. $\sum_{g \in G_{+, s}} \mu_g^h|((\delta - \mu_g)/\sigma_v|^k \phi((\delta - \mu_g)/\sigma_v)/N = O_p(G/N)$ and $\sum_{g \in G_{+, s}} \mu_g^h \Phi((\delta - \mu_g)/\sigma_v)/N = O_p(G/N)$ for $h = 1, 2$;

4. $\sum_{g \in G_{+, s}} \mu_g^h|((\delta - \mu_g)/\sigma_v|^k \phi((\delta - \mu_g)/\sigma_v) = O_p(G^2/N^2) = o_p(G/N)$, $\sum_{g \in G_{+, s}} \mu_g^h \Phi((\delta - \mu_g)/\sigma_v) = O_p(G^2/N^2) = o_p(G/N)$, $\sum_{g \in G_{+, s}} \mu_g^h|((\delta - \mu_g)/\sigma_v|^k \phi((\delta - \mu_g)/\sigma_v) = O_p((G/N)^{3/2}) = o_p(G/N)$ and $\sum_{g \in G_{+, s}} \mu_g^h \Phi((\delta - \mu_g)/\sigma_v) = O_p((G/N)^{3/2}) = o_p(G/N)$.

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

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

For the second statement, notice that to show the first inequality we only need to show that $E[|\delta - \mu_g|^k e^{-(\delta - \mu_g)^2}] \lesssim \frac{1}{\eta^n}$. For any non-negative integer $k$, we have

$$E[|\delta - \mu_g|^k e^{-(\delta - \mu_g)^2}] \leq E[e^{k(|\delta - \mu_g| - 1) - (\delta - \mu_g)^2}] = e^{k^2/4 - k} \cdot E[e^{-(|\delta - \mu_g| - k/2)^2}],$$

with $E[e^{-(|\delta - \mu_g| - k/2)^2}]$ bounded by

$$E[e^{-(|\delta - \mu_g| - k/2)^2}] \leq \zeta_g P(e^{-(|\delta - \mu_g| - k/2)^2} \leq \zeta_g) + E[e^{-(|\delta - \mu_g| - k/2)^2} > \zeta_g] \leq \zeta_g + P(e^{-(|\delta - \mu_g| - k/2)^2} > \zeta_g) \leq \zeta_g + P(e^{-(|\delta - \mu_g| - k/2)^2} > \zeta_g, \delta < \sqrt{1 - \eta \mu_g}) + P(e^{-(|\delta - \mu_g| - k/2)^2} > \zeta_g, \delta \geq \sqrt{1 - \eta \mu_g}) \leq \zeta_g + P(e^{-(1 - \sqrt{1 - \eta \mu_g})^2 \mu_g^2/2} > \zeta_g) + P(\delta \geq \sqrt{1 - \eta \mu_g}),$$

39
where the last inequality holds for large enough \( n_g \) as both \( k \) and \( \eta \) are fixed. Set \( \zeta_g = e^{-n_g k_g (1-\sqrt{1-\eta})^2 \rho_g^2 / 8} \), we know that \( \zeta_g \leq e^{-n_g E(1-\sqrt{1-\eta})^2 \rho_g^2 / 8} \lesssim \frac{1}{n_g} \), and \( P(e^{-n_g E(1-\sqrt{1-\eta})^2 \rho_g^2 / 2} > \zeta_g) = P(Z_g' z_g < k_g n_g / 4) \lesssim \frac{1}{n_g^2} \). The inequality is then proven as \( P(\delta \geq 1 - \eta \cdot \mu_g) \lesssim \frac{1}{n_g} \) as already shown at the end of the proof for Lemma A3.

For the second inequality in the second statement, notice that for the same \( \eta \) as above,
\[
E\left[ \Phi\left( \frac{\delta - \mu_g}{\sigma_v} \right) \right] = E\left[ \Phi\left( \frac{\delta - \mu_g}{\sigma_v} \right) 1(\delta \leq 1 - \eta \cdot \mu_g) \right] + E\left[ \Phi\left( \frac{\delta - \mu_g}{\sigma_v} \right) 1(\delta > 1 - \eta \cdot \mu_g) \right]
\]
\[
\leq E\left[ \Phi\left( \frac{1}{\sigma_v} \right) \right] + P(\delta > 1 - \eta \cdot \mu_g)
\]
\[
\leq E\left[ \Phi\left( \frac{1}{\sigma_v} \right) \right] + P(\mu_g > \sqrt{n_g \rho_g \sqrt{k_g / 2}}) + P(\mu_g \leq \sqrt{n_g \rho_g \sqrt{k_g / 2}}) + P(\delta > 1 - \eta \cdot \mu_g)
\]
\[
\leq \Phi\left( \frac{1}{\sigma_v} \right) + P(Z_g' z_g / n_g \leq k_g / 2) + P(\delta > 1 - \eta \cdot \mu_g).
\]

The inequality is hence proven as \( \Phi(-d) \leq \phi(d) / d \lesssim \frac{1}{n_g^2} \) with \( d = (1 - \sqrt{1 - \eta}) \sqrt{n_g \rho_g \sqrt{k_g / 2}} / \sigma_v \) for \( n_g \) large. \( P(Z_g' z_g / n_g \leq k_g / 2) \lesssim \frac{1}{n_g^2} \) following from Lemma A2, and \( P(\delta > 1 - \eta \cdot \mu_g) \lesssim \frac{1}{n_g} \) as shown in the proof of Lemma A3.

The third statement is apparent given the first statement and the fact that by MI
\[
\sum_{g \in G_{+},w} \mu_g / N = O_p \left( \bar{\rho} \sum_{g \in G_{+},w} E\left[ \sqrt{Z_g' z_g / n_g} / N \right] \right) = O_p \left( \bar{\rho} \sum_{g \in G_{+},w} \sqrt{E\left[ Z_g' z_g / n_g \right] / N} \right) = O_p(\bar{\rho}^2 \sqrt{kG} / N) = O_p(G / N).
\]

To prove the last statement, notice that by the second statement, it is clear that if we can show the case with \( k = 0 \), the same result with all other positive \( k \) will follow. For simplicity, use \( \phi_{g,\delta} \) to denote \( \phi((\delta - \mu_g) / \sigma_v) \). Then we have
\[
E[\mu_g \phi_{g,\delta}] = E[\mu_g \phi_{g,\delta}^{1/2} \cdot \phi_{g,\delta}^{1/2}] \leq \sqrt{E[\mu_g^2 \phi_{g,\delta}]} \sqrt{E[\phi_{g,\delta}]} \lesssim E[\phi_{g,\delta}]^{1/4} E[\mu_g^2]^{1/4} E[\phi_{g,\delta}]^{1/2} \]
\[
\lesssim \bar{\rho} E[Z_g^2]^{1/4} E[\phi_{g,\delta}]^{3/4} \lesssim \bar{\rho} E\left[ (Z_g' z_g) \right]^{1/4} E[\phi_{g,\delta}]^{3/4} \lesssim 1 / n_g,
\]

where the last inequality follows from the third statement of the lemma and the moment condition required in Assumption 1. Similarly, \( E[\mu_g^2 \phi_{g,\delta}] \lesssim 1 / \sqrt{n_g}, E[\mu_g \Phi_{g,\delta}] \lesssim 1 / n_g, \) and \( E[\mu_g^2 \Phi_{g,\delta}] \lesssim 1 / \sqrt{n_g} \). The last statement is therefore proven. \( \square \)
Lemma A5. Under assumption 1, if $G \log G/N \to 0$ as $G,N \to \infty$, with probability goes to one,

$$\min_{g \in G_{+,s}} \mu^2_g \gtrsim N/G$$

Proof. Note that we have $Z_{ig} \equiv \bar{Z}_{ig} - \hat{\gamma} X_{ig}$ where $\hat{\gamma}$ is the OLS estimator for the coefficient $\gamma$ of regressing $\bar{Z}_g$ on $X_g$, therefore for all $g$ we have,

$$\sum_i \rho^2_g Z^2_{ig} = \rho^2_g \sum_i (\bar{Z}_{ig} - \gamma X_{ig} + \gamma X_{ig} - \hat{\gamma} X_{ig})^2$$

$$\geq \rho^2_g \sum_i (\bar{Z}_{ig} - \gamma X_{ig})^2 + 2(\gamma - \hat{\gamma}) \sum_i (\bar{Z}_{ig} - \gamma X_{ig}) X_{ig}$$

Since $(\bar{Z}_{ig} - \gamma X_{ig})^2$ are independent within group $g$ and are non-negative, we can use the one-sided Bernstein inequality to bound its lower tail:

$$P\left(\rho^2_g \sum_i (\bar{Z}_{ig} - \gamma X_{ig})^2 - \mathbb{E}[(\bar{Z}_{ig} - \gamma X_{ig})^2] \leq -n_g \delta\right) \leq \exp\left(-\frac{n_g \delta^2}{2\rho^4_g \mathbb{E}[(\bar{Z}_{ig} - \gamma X_{ig})^4]}\right)$$

for any $\delta > 0$. Under assumption 1, we have $\Delta_Z \leq \mathbb{E}[(\bar{Z}_{ig} - \gamma X_{ig})^4] \leq \tilde{\Delta}_Z$ and $\tilde{\zeta}_Z \leq \mathbb{E}[(\bar{Z}_{ig} - \gamma X_{ig})^2] \leq \tilde{\zeta}_Z$. Pick $\delta = \mathbb{E}[(\bar{Z}_{ig} - \gamma X_{ig})^2]/2$, then

$$P\left(\rho^2_g \sum_i (\bar{Z}_{ig} - \gamma X_{ig})^2 \leq \frac{c \zeta_Z N}{G}\right) \leq P\left(\rho^2_g \sum_i (\bar{Z}_{ig} - \gamma X_{ig})^2 \leq n_g \mathbb{E}[(\bar{Z}_{ig} - \gamma X_{ig})^2]/2\right)$$

$$\leq \exp\left(-\frac{c \zeta^2_Z N}{8\rho^4 \Delta_Z}\right)$$

hence

$$P\left(\min_{g \in G_{+,s}} \rho^2_g (\bar{Z}_{ig} - \gamma X_{ig})^2 \leq \frac{c \zeta_Z N}{G}\right) = P\left(\exists g \in G_{+,s}, \mu^2_g \leq \frac{c \zeta_Z N}{G}\right) \leq \sum_{g \in G_{+,s}} P\left(\mu^2_g \leq \frac{c \zeta_Z N}{G}\right)$$

$$\leq G \exp\left(-\frac{c \zeta^2_Z N}{8\rho^4 \Delta_Z}\right) = \exp\left(\log G - \frac{c \zeta^2_Z N}{8\rho^4 \Delta_Z}\right)$$

$$= \exp\left(\frac{N}{G} \left(\frac{G \log G}{N} - \frac{c \zeta^2_Z N}{8\rho^4 \Delta_Z}\right)\right) \to 0$$

Now combined with the fact that that $\hat{\gamma} - \gamma = O_p(\frac{1}{N/G}) = o_p(1)$ hence $\rho^2_g \sum_i (\bar{Z}_{ig} - \gamma X_{ig})^2$ dominates $2(\gamma - \hat{\gamma}) \sum_i (\bar{Z}_{ig} - \gamma X_{ig}) X_{ig}$, the Lemma is proven. 

\[\square\]
Proof of Theorem 1

Proof. We only prove for the case where $G$ goes to infinity together with $N$. The case for fixed $G$ follows a similar but much simpler argument. We will also repeatedly use the property that truncated mean monotonically increases with the truncation cutoff since
\[
\frac{\partial}{\partial y} E[X|X > y] = \frac{f_X(y)}{1-F_X(y)} (E[X|X > y] - y) \geq 0 \text{ for any random variable } X \text{ and truncation cutoff } y.
\]
Further, although Assumption 1 assumes both homoskedasticity and one-sided first stage relationship, we prove the 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. Let $\sigma^2_{g,v} = E[v_{ig}^2]$ with $0 < \sigma_v \leq \sigma_{g,v} \leq \bar{\sigma}_v < \infty$ for all $g = 1, ..., G$. And let $\mathcal{G}_- = \{g : \rho_g < 0\}$ denote the set of groups with negative first-stage effects. Further define $G_- = |\mathcal{G}_-| = G - G_0 - G_{+,a} - G_{+,w}$.

Since
\[
\sqrt{N/G} \cdot E\left[ \sum_{g=1}^G Z'_g u_g 1(t_g > c_g) \frac{1}{\sum_{g=1}^G n_g 1(t_g > c_g)} \right] 
\]
\[
\geq \frac{1}{\sqrt{NG}} E\left[ \sum_{g=1}^G Z'_g u_g 1(t_g > c_g) \right] - \frac{1}{\sqrt{NG}} E\left[ \sum_{g=1}^G Z'_g u_g 1(t_g > c_g) \left( \sum_{g=1}^G 1(t_g > c_g) = 0 \right) \right]
\]
\[
= \frac{1}{\sqrt{NG}} E\left[ \sum_{g=1}^G Z'_g u_g 1(t_g > c_g) \right],
\]
it suffices to show that there exists some positive constant $a^*$ such that
\[
\frac{1}{G} E\left[ \sum_{g=1}^G \frac{Z'_g v_g}{\sqrt{n_g}} 1(t_g > c_g) \right] \geq a^* + o(1).
\]
Decompose the left hand side, we have that

\[
\frac{1}{G} E \left[ \sum_{g=1}^{G} \frac{Z'_g v_g}{\sqrt{n_g}} 1(t_g > c_g) \right] = \frac{1}{G} E \left[ \sum_{g \in \mathcal{V}_{+}} \frac{Z'_g v_g}{\sqrt{n_g}} 1(t_g > c_g) \right] + \frac{1}{G} E \left[ \sum_{g \in \mathcal{V}_{0}} \frac{Z'_g v_g}{\sqrt{n_g}} 1(t_g > c_g) \right] + \frac{1}{G} E \left[ \sum_{g \in \mathcal{V}_{-}} \frac{Z'_g v_g}{\sqrt{n_g}} 1(t_g > c_g) \right] = A + B + D + E.
\]

Term A is non-negative as

\[
A = \frac{1}{G} \sum_{g \in \mathcal{V}_{+}} E \left[ \frac{Z'_g v_g}{\sqrt{n_g}} \frac{Z_g v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2} - \sqrt{n_g} H_{g,2}^2 \right] P[t_g > c_g]
\]

Similary we can show that term E is non-negative.

Let \( \delta_c = \inf_g c_g \) and \( \Delta_c = \sup_g c_g \). Since \( 0 < \alpha < 0.5 \), \( \delta_c > 0 \) and \( \Delta_c < \infty \). Let \( I_g = \{|H_{g,1} - \sigma^2_{g,v}| \leq \frac{1}{2} \sigma^2_{g,v}, |H_{g,2}^2 - k_g| \leq \frac{1}{2} k_g\} \). Applying Lemma A2 we have that \( \frac{1}{G} \sum_g P(I_g) = O(G^2/N^2) = o(1) \). Therefore,

\[
B = \frac{1}{G} E \left[ \sum_{g \in \mathcal{V}_{0}} \frac{Z'_g v_g}{\sqrt{n_g}} 1(t_g > c_g) \right] = \frac{1}{G} E \left[ \sum_{g \in \mathcal{V}_{0}} \frac{Z'_g v_g}{\sqrt{n_g}} 1\left( \frac{Z'_g v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2} \right) \right]
\]

\[
\geq \frac{1}{G} \sum_{g \in \mathcal{V}_{0}} c_g E \left[ H_{g,1} H_{g,2} 1\left( \frac{Z'_g v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2} \right) \right]
\]

\[
\geq \frac{1}{G} \sum_{g \in \mathcal{V}_{0}} c_g E \left[ H_{g,1} H_{g,2} 1(I_g) \left( \frac{Z'_g v_g}{\sqrt{n_g}} > c_g H_{g,1} H_{g,2} \right) \right]
\]

\[
\geq \frac{1}{G} \sum_{g \in \mathcal{V}_{0}} c_g E \left[ H_{g,1} H_{g,2} 1(I_g) \left( \frac{Z'_g v_g}{\sqrt{n_g}} > c_g \frac{3}{2} \sigma_{g,v} \sqrt{k_g} \right) \right]
\]

\[
\geq \frac{1}{2} \sigma^2_{g,v} \sqrt{k_g} \delta_c \sum_{g \in \mathcal{V}_{0}} \left( P\left[ \frac{Z'_g v_g}{\sqrt{n_g k_g \sigma_{g,v}}} > \frac{3}{2} c_g \right] - P[I_g] \right)
\]

\[
= \frac{1}{2} \sigma^2_{g,v} \sqrt{k_g} \delta_c \sum_{g \in \mathcal{V}_{0}} \left[ \frac{Z'_g v_g}{\sqrt{n_g k_g \sigma_{g,v}}} > \frac{3}{2} c_g \right] + o(1).
\]
Let $\eta_g$ be the resulted vector of error terms from a linear projection of $\tilde{Z}_g$ on $X_g$, then $Z'_g Z_g/n_g \overset{p}{\to} E[\eta_g^2] = k_g$. Then

$$\frac{1}{G} \sum_{g \in G_0} P \left[ \frac{Z'_g v_g}{\sqrt{n_g k_g \sigma_g v}} > \frac{3}{2} c_g \right] = \frac{1}{G} \sum_{g \in G_0} \left( 1 - P \left[ \frac{Z'_g v_g}{\sqrt{n_g k_g \sigma_g v}} \leq \frac{3}{2} c_g \right] \right)$$

$$\geq G_0/G - \frac{1}{G} \sum_{g \in G_0} P \left[ \frac{\eta'_g v_g}{\sqrt{n_g k_g \sigma_g v}} \leq \frac{3}{2} c_g + \frac{\eta'_g P_{X_g} v_g}{\sqrt{n_g k_g \sigma_g v}} \right]$$

$$\geq G_0/G - \frac{1}{G} \sum_{g \in G_0} P \left[ \frac{\eta'_g v_g}{\sqrt{n_g k_g \sigma_g v}} \leq \frac{3}{2} c_g + \frac{\eta'_g P_{X_g} v_g}{\sqrt{n_g k_g \sigma_g v}}, \frac{\eta'_g P_{X_g} v_g}{\sqrt{n_g k_g \sigma_g v}} \leq c_g/2 \right]$$

$$- \frac{1}{G} \sum_{g \in G_0} P \left[ \frac{\eta'_g P_{X_g} v_g}{\sqrt{n_g k_g \sigma_g v}} > c_g/2 \right]$$

$$\geq G_0/G - \frac{1}{G} \sum_{g \in G_0} P \left[ \frac{\eta'_g v_g}{\sqrt{n_g k_g \sigma_g v}} \leq 2c_g \right] - \frac{1}{G} \sum_{g \in G_0} \left[ \eta'_g P_{X_g} v_g \right] \sqrt{n_g k_g \sigma_g v} c_g/2$$

$$\geq G_0/G - \frac{1}{G} \sum_{g \in G_0} P \left[ \frac{\eta'_g v_g}{\sqrt{n_g k_g \sigma_g v}} \leq 2c_g \right] - \frac{1}{G} \sum_{g \in G_0} \sqrt{E \left[ (\eta'_g P_{X_g} v_g)^2 \right]} \sqrt{n_g k_g \sigma_g v c_g/2}$$

$$\geq G_0/G - \frac{1}{G} \sum_{g \in G_0} P \left[ \frac{\eta'_g v_g}{\sqrt{n_g k_g \sigma_g v}} \leq 2c_g \right] - \frac{1}{G} \sum_{g \in G_0} \sqrt{E \left[ (\eta'_g P_{X_g} v_g)^2 \right]} \sqrt{n_g k_g \sigma_g v c_g/2}$$

$$= \sum_{g \in G_0} \Phi(-2c_g) + O \left( 1/\sqrt{N/G} \right) + O(1/\sqrt{N/G})$$

$$\geq G_0/G \Phi(-2\Delta_c) + o(1)$$

where the third last step is by the Markov inequality, the second last step is by the Cauchy-Schwarz inequality, and the last step is by the Berry-Esseen theorem as $E[|\eta_g v_g|^3]$ is bounded by a universal constant for all $g = 1, ..., G$ following Assumption 1, and the fact that $E \left[ (\eta'_g P_{X_g} v_g)^2 \right]$ is bounded by a universal constant across $g = 1, ..., G$ following a similar result shown in the proof of Lemma A2. Therefore, we know that

$$B \geq \frac{G_0}{G} \cdot \frac{1}{2} \sqrt{\sigma_x \delta_c} \Phi(-2\Delta_c) + o(1).$$
Similarly, for term $D$, we have

\[
D = \frac{1}{G} \sum_{g \in G_{+},w} E \left[ \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} \left( \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > c_{g}H_{g,1}H_{g,2} - a_{g}H_{g,2}^{2} \right) 1(\hat{g}\neq g) \right] P \left[ \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > c_{g}H_{g,1}H_{g,2} - a_{g}H_{g,2}^{2} \right]
\]

\[
\geq \frac{1}{G} \sum_{g \in G_{+},w} E \left[ \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} \left( \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > -a_{g}H_{g,2}^{2} \right) P \left( \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > c_{g}H_{g,1}H_{g,2} \right) \right]
\]

\[
\geq \frac{1}{G} \sum_{g \in G_{+},w} E \left[ \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} \left( \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > -a_{g}H_{g,2}^{2} \right) P \left[ \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > c_{g}H_{g,1}H_{g,2} \right] \right]
\]

\[
= \frac{1}{G} \sum_{g \in G_{+},w} E \left[ \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} \left( \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > -a_{g}H_{g,2}^{2} \right) 1(\hat{g}\neq g) \right] P \left( \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > c_{g}H_{g,1}H_{g,2}, \hat{g}\neq g \right)
\]

\[
\geq \rho k \frac{1}{2} G \sum_{g \in G_{+},w} E \left[ 1 \left( \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} \leq -\frac{\bar{p}k_{g}}{2} \right) 1(\hat{g}\neq g) \right] P \left( \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > c_{g}3\sigma_{g,v}\sqrt{k_{g}} \right)
\]

\[
\geq \rho k \frac{1}{2} G \sum_{g \in G_{+},w} \left( P \left[ \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} \leq -\frac{\bar{p}k_{g}}{2} \right] - P(\hat{g}) \right) \left( P \left[ \frac{Z_{g}'v_{g}}{\sqrt{n_{g}}} > c_{g}3\sigma_{g,v}\sqrt{k_{g}} \right] \right) - P(\hat{g})
\]

\[
= \frac{G_{+},w}{G} \cdot \frac{\rho k}{2} \left( -2\frac{\bar{p}\sigma_{v}}{2} \right) \Phi \left( -2\Delta_{v} \right) + o(1).
\]

Since under Assumption 1, $(G_{0} + G_{+},w)/G \rightarrow 1 - b$ which is bounded away from zero. Putting together results stated above, the theorem is proven. As discussed in the main text, the theorem holds under the more general setting with group-specific first-stage error term variance and unrestricted signs in first stage effects than is required in Assumption 1. In the rest of the proofs, we will follow the restrictions imposed in Assumption 1 for notation simplicity.

**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
\[
\sum_{g=1}^{G} \hat{\rho}_g Z'_{g} W_g 1(\hat{\mu}_g > \delta) / N = \sum_{g=1}^{G} (\rho_g + (Z'_{g} Z_g)^{-1} Z'_{g} v_g) (\rho_g Z'_{g} Z_g + Z'_{g} v_g) 1(\hat{\mu}_g > \delta) / N \\
= \frac{1}{N} \sum_{g \in \mathcal{G}_{+}, w} \rho_g Z'_{g} Z_g 1(\hat{\mu}_g > \delta) + \frac{1}{N} \sum_{g \in \mathcal{G}_{+}, s} \rho_g Z'_{g} Z_g 1(\hat{\mu}_g > \delta) \\
+ 2 \frac{1}{N} \sum_{g=1}^{G} \rho_g Z'_{g} v_g 1(\hat{\mu}_g > \delta) + \frac{1}{N} \sum_{g=1}^{G} (Z'_{g} Z_g)^{-1} (Z'_{g} v_g)^2 1(\hat{\mu}_g > \delta) \\
= A_I + A_{II} + B + C.
\]

First, \(A_I, B, \) and \(C\) are \(o_p(1)\) by Markov’s inequality and the fact that
\[
E[|A_I|] \leq \frac{1}{N} \sum_{g \in \mathcal{G}_{+}, w} a_g^2 E[Z'_{g} Z_g / n_g] \leq \bar{k} \rho \cdot G_{+}, w / N \to 0,
\]
\[
E[|B|] \leq \frac{1}{N} \sum_{g} \rho_g E[|Z'_{g} v_g|] \leq \frac{1}{N} \sum_{g} \rho_g \sqrt{E[(Z'_{g} v_g)^2]} = \frac{1}{N} \sum_{g} \rho_g \sqrt{\sigma^2_{g,v} n_g k_g} \leq \sqrt{\sigma^2_{k} \bar{c} G / N} \to 0.
\]
\[
E[|C|] \leq E \left[ \frac{1}{N} \sum_{g} (Z'_{g} Z_g)^{-1} (Z'_{g} v_g)^2 \right] = \frac{1}{N} \sum_{g} \sigma^2_{g,v} \leq \frac{G^2}{N} \sigma^2_v \to 0.
\]

Then, we prove that \(A_{II} = k'' + o_p(1)\). Since it is clear \(\frac{1}{N} \sum_{g \in \mathcal{G}_{+}, s} \rho_g^2 Z'_{g} Z_g \rightarrow_{p} k''\) by the Law of Large Numbers for independent and non-identically distributed random variable [Corollary 3.9, White (2000)] and Assumption 1.3, 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 \(\delta\) range in Assumption 2, there exists a small positive constant
$\eta \in (0, 1)$ such that $\delta \leq \rho \sqrt{\kappa c} (1 - \eta) \sqrt{N/G}$,

\[
E \left( \left| \frac{1}{N} \sum_{g \in G_{+, s}} \rho_g^2 Z_g Z_g(1(\hat{\mu}_g > \delta) - 1) \right| \right) \leq \frac{1}{N} \sum_{g \in G_{+, s}} \rho_g^2 E \left[ Z_g Z_g 1(\hat{\mu}_g \leq \delta) \right]
\]
\[
\leq \frac{1}{N} \sum_{g \in G_{+, s}} \rho_g^2 n_g \sqrt{E[(Z_g' Z_g/n_g)^2]} \sqrt{P(\hat{\mu}_g \leq \delta)}
\]
\[
\leq \frac{1}{N} \sum_{g \in G_{+, s}} \rho_g^2 n_g \sqrt{E[\tilde{Z}_g' \tilde{Z}_g/n_g]^2]} \sqrt{P(\tilde{\mu}_g \leq \rho \sqrt{k c} (1 - \eta) \sqrt{N/G})}
\]
\[
\leq \frac{1}{N} \sum_{g \in G_{+, s}} \bar{\rho}^2 \sqrt{E[(\tilde{Z}_g' \tilde{Z}_g/n_g)^2]} \sqrt{P(\tilde{\mu}_g \leq \rho \sqrt{k c} (1 - \eta))}
\]
\[
\rightarrow 0.
\]

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

Then we consider

\[
\frac{1}{\sqrt{N}} \sum_{g=1}^{G} \rho_g Z_g' u_g 1(\hat{\mu}_g > \delta)
\]
\[
= \frac{1}{\sqrt{N}} \sum_{g} \rho_g Z_g' u_g - \frac{1}{\sqrt{N}} \sum_{g} \rho_g Z_g' u_g 1(\hat{\mu}_g \leq \delta) + \frac{1}{\sqrt{N}} \sum_{g} (Z_g' v_g)(Z_g' Z_g)^{-1} Z_g' u_g 1(\hat{\mu}_g > \delta)
\]
\[
= F_1 + F_2 + F_3
\]

Under Assumption 1.3, $F_1 \Rightarrow N(0, \sigma_u^2 k''')$ by the Liapounov Central Limit theorem [Theorem 5.10 in White (2000)].
In addition, $F_3$ and $F_2$ are $o_p(1)$ by the Markov’s inequality and the fact that

$$E[|F_3|] \leq E\left[\frac{1}{\sqrt{N}} \sum_g Z_g^v Z_g^u \frac{Z_g^t Z_g}{\sqrt{Z_g^t Z_g}}\right] \leq \frac{1}{\sqrt{N}} \sum_g \sqrt{E\left[\left(\frac{Z_g^t Z_g}{\sqrt{Z_g^t Z_g}}\right)^2\right]} \sqrt{E\left[\left(\frac{Z_g^t Z_g}{\sqrt{Z_g^t Z_g}}\right)^2\right]} \leq \frac{G}{\sqrt{N}} \sigma u \sigma_v \rightarrow 0,$$

$$E[|F_2|] \leq \frac{1}{\sqrt{N}} \sum_{g \in \mathcal{G}^+,s} E[|\rho_g Z_g^u 1(\hat{\mu}_g \leq \delta)|] + \frac{1}{\sqrt{N}} \sum_{g \in \mathcal{G}^+,w} E[|\rho_g Z_g^u 1(\hat{\mu}_g \leq \delta)|] \leq \frac{1}{\sqrt{N}} \sum_{g \in \mathcal{G}^+,s} \rho_g \sqrt{E\left[\left(Z_g^u / \sqrt{n_g}\right)^2\right]} \sqrt{P(\hat{\mu}_g \leq \delta)} + \frac{1}{\sqrt{N}} \sum_{g \in \mathcal{G}^+,w} \rho_g \sqrt{E\left[\left(Z_g^u / \sqrt{n_g}\right)^2\right]} \leq O(G/\sqrt{N} \cdot N/G \cdot 1/(N/G)) + O(G/\sqrt{N}) = o(1).$$

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

Combining all pieces and apply the Slutsky’s Theorem we know that with $\tilde{H} = k''$,

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

\[\Box\]

**Proof for Lemma 2**

**Proof.** First we consider the asymptotic property of $\hat{\beta}^a(\delta)$, the split-sample estimator with sample $a$. Notice that we have

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

48
Note that the denominator follows that
\[
\frac{1}{N^a} \sum_g (W^b_g)' Z^b_g (Z^b_g)' Z^g_a - 1 (Z^a_g)' W^a_g 1(\hat{\mu}_g > \delta)
\]
\[
= \frac{1}{N^a} \sum_g (\rho_g + ((Z^b_g)' Z^b_g - 1 (Z^b_g)' \nu^b_g) (\rho_g (Z^a_g)' Z^a_g + (Z^a_g)' \nu^a_g) 1(\hat{\mu}_g > \delta)
\]
\[
= \frac{1}{N^a} \sum_g \rho_g^2 (Z^a_g)' Z^a_g 1(\hat{\mu}_g > \delta) + \frac{1}{N^a} \sum_g \rho_g (Z^a_g)' \nu^a_g 1(\hat{\mu}_g > \delta)
\]
\[
+ \frac{1}{N^a} \sum_g \rho_g (Z^a_g)' (Z^a_g)' (Z^b_g)' \nu^b_g 1(\hat{\mu}_g > \delta) + \frac{1}{N^a} \sum_g (Z^a_g)' Z^g_a - 1 (Z^b_g)' \nu^b_g (Z^a_g)' \nu^a_g 1(\hat{\mu}_g > \delta)
\]
\[
= \frac{1}{N^a} \sum_g \rho_g^2 Z^a_g Z^a_g 1(\hat{\mu}_g > \delta) + o_p(1),
\]
\[
k'' + o_p(1).
\]
where the second last line uses the fact that subsamples \(a\) and \(b\) are independent of each other and the last line could be derived using similar argument as in Lemma 1 for showing the probability limit of \(\frac{1}{N^a} \sum_g \rho_g^2 Z^a_g Z^a_g 1(\hat{\mu}_g > \delta)\).

The numerator
\[
\sum_g (W^b_g)' Z^b_g (Z^b_g)' Z^g_a - 1 (Z^a_g)' W^a_g 1(\hat{\mu}_g > \delta) / \sqrt{N^a}
\]
\[
= \frac{1}{\sqrt{N^a}} \sum_g \rho_g (Z^a_g)' u^a_g - \frac{1}{\sqrt{N^a}} \sum_g \rho_g (Z^a_g)' u^a_g 1(\hat{\mu}_g \leq \delta) + \frac{1}{\sqrt{N^a}} ((Z^b_g)' Z^g_a - 1 (Z^b_g)' \nu^b_g (Z^a_g)' u^a_g 1(\hat{\mu}_g > \delta)
\]
\[
= \frac{1}{\sqrt{N^a}} \sum_g \rho_g (Z^a_g)' u^a_g + o_p(1),
\]
where the last equality again uses the fact that subsamples are independent of each other.

Similar derivations also hold for the split-sample estimator in sample \(b\). Putting together,
\[
\sqrt{N} (\hat{\beta}_{sssel,int}(\delta) - \beta) = \frac{1}{2} \sqrt{N/N^a} \sqrt{N^a} (\hat{\beta}^a(\delta) - \beta) + \frac{1}{2} \sqrt{N/N^b} \sqrt{N^b} (\hat{\beta}^b(\delta) - \beta)
\]
\[
= \frac{1}{\sqrt{2}} (h^a + h^b) / k'' + o_p(1)
\]
where \(h^a = \frac{1}{\sqrt{N^a}} \sum_g \rho_g Z^a_g u^a_g \) and \(h^b = \frac{1}{\sqrt{N^b}} \sum_g \rho_g^2 Z^b_g u^b_g\). Since \(h^a\) and \(h^b\) are independent, we have that \((h^a, h^b)'\) converges jointly to \(N((0, 0)', \sigma^2_a k'' I_2)\). Therefore,
\[
\sqrt{N} (\hat{\beta}_{sssel,int}(\delta) - \beta) \Rightarrow N(0, \sigma^2_a H^{-1})
\]
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 $\delta$, the decomposition in the following Lemma depends on $\delta$.

**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

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

such that

1) $h = O_p(1)$, $H = O_p(1)$, 2) $\rho_{\delta,N} \equiv S(\delta) = o_p(1)$, 3) $T^h = o_p(1)$,
4) $(T^H)^2 = o_p(\rho_{\delta,N})$, 5) $(T^H)(T^h) = o_p(\rho_{\delta,N})$, 6) $Z^H = o_p(\rho_{\delta,N})$,
7) $E[\hat{A}(\delta)|\tilde{Z}, \tilde{X}] = \sigma_u^2H + S(\delta)H^2 + o_p(\rho_{\delta,N})$.
8) $Z^A(\delta) = o_p(\rho_{\delta,N})$

then

$N(\hat{\beta} - \beta_0)^2 = \hat{Q}(\delta) + \hat{r}(\delta)$

$E[\hat{Q}(\delta)|\tilde{Z}, \tilde{X}] = \sigma_u^2H^{-1} + S(\delta) + T(\delta)$

$\frac{(\hat{r}(\delta) + T(\delta))}{S(\delta)} = o_p(1)$ 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 in the expectation in 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 \( \hat{h}_\delta = \frac{W_P e}{\sqrt{N}} \) and \( \hat{H}_\delta = \frac{W_P W}{N} \) and \( P_\delta \) is a block diagonal matrix consisting matrices \( Z_g(Z_g'Z_g)^{-1}Z_g'1(\hat{\mu}_g > \delta) \) on its diagonals. Let \( f = [\rho_1Z_1' \rho_2Z_2' \ldots \rho_G Z_G']', \hat{h}_\delta = h + T_1^h + T_2^h, \) and \( \hat{H}_\delta = H + T_1^H + T_2^H + T_3^H + Z^H \) with

\[
h = f' u / \sqrt{N} = \sum_g \rho_g Z_g' u_g / \sqrt{N}; \quad H = \sum_{g \in \mathcal{U}^+,s} \rho_g^2 Z_g' Z_g / N = \sum_{g \in \mathcal{U}^+,s} \mu_g^2 / N;
\]

\[
T_1^h = -f' (I - P_\delta) u / \sqrt{N} = -\sum_g \rho_g 1(\hat{\mu}_g < \delta) Z_g' u_g / \sqrt{N};
\]

\[
T_2^h = v' P_\delta u / \sqrt{N} = \sum_g v'_g Z_g (Z_g'Z_g)^{-1}Z_g' u_g 1(\hat{\mu}_g \geq \delta) / \sqrt{N}; \quad T_1^H = f' f / N - H = \sum_{g \in \mathcal{V}^+,w} \mu_g^2 / N;
\]

\[
T_2^H = -f' (I - P_\delta) f / N; \quad T_3^H = (v' f + f' v) / N; \quad Z^H = (v' P_\delta v - v' (I - P_\delta) f - f' (I - P_\delta) v) / N.
\]

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

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:

\[
E[1(\hat{\mu}_g > \delta)] = 1 - \Phi_{g,\delta}
\]

\[
E[1(\hat{\mu}_g > \delta) Z_g' v_g] = \sigma_v \phi_{g,\delta} \sqrt{Z_g' Z_g}
\]

\[
E[1(\hat{\mu}_g > \delta) (Z_g' v_g)^2] = \sigma_v^2 (1 - \Phi_{g,\delta} + t_{g,\delta} \phi_{g,\delta}) Z_g' Z_g
\]

\[
E[1(\hat{\mu}_g > \delta) (Z_g' v_g)^3] = \sigma_v^3 \phi_{g,\delta} (t_{g,\delta}^2 + 2) (Z_g' Z_g)^{3/2}
\]

\[
E[1(\hat{\mu}_g > \delta) (Z_g' v_g)^4] = \sigma_v^4 \phi_{g,\delta} (t_{g,\delta}^3 + 3t_{g,\delta}) (Z_g' Z_g)^2 + 3(1 - \Phi_{g,\delta})(Z_g' Z_g)^2
\]

\[
E[1(\hat{\mu}_g > \delta) Z_g' u_g] = \frac{\sigma_u \sigma_w}{\sigma_v} \phi_{g,\delta} \sqrt{Z_g' Z_g}
\]

\[
E[1(\hat{\mu}_g > \delta)(Z_g' u_g)(Z_g' v_g)] = \frac{\sigma_u \sigma_w}{\sigma_v^3} E[1(\hat{\mu}_g > \delta)(Z_g' v_g)^2] = \sigma_u (1 - \Phi_{g,\delta} + t_{g,\delta} \phi_{g,\delta}) Z_g' Z_g
\]

\[
E[1(\hat{\mu}_g > \delta)(Z_g' u_g)^2] = \frac{\sigma_u^2 \sigma_w^2}{\sigma_v^4} E[1(\hat{\mu}_g > \delta)(Z_g' v_g)^2] + (\sigma_u^2 - \frac{\sigma_u^2 \sigma_w^2}{\sigma_v^2}) E[1(\hat{\mu}_g > \delta)] Z_g' Z_g
\]

\[
= \left( \frac{\sigma_u^2}{\sigma_v^2} (1 - \Phi_{g,\delta}) + \frac{\sigma_u^2 \sigma_w^2}{\sigma_v^4} t_{g,\delta} \phi_{g,\delta} \right) Z_g' Z_g
\]

**Asymptotic MSE for the Select-and-Interact Estimator**

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

1) \( h = O_p(1), \quad H = O_p(1), \quad \rho_{\delta,N} = S_{\text{sel},\text{int}}(\delta) = o_p(1), \) 
2) \( T_{1h} + T_{2h} = o_p(1), \) 
3) \( (T_{1H} + T_{2H} + T_{3H})^2 = o_p(\rho_{\delta,N}^3), \) 
4) \( (T_{1H} + T_{2H} + T_{3H})(T_{1h} + T_{2h}) = o_p(\rho_{\delta,N}), \) 
5) \( Z_H = o_p(\rho_{\delta,N}), \) 
6) \( \sigma_u^2 H + H^2 S_{\text{sel},\text{int}}(\delta) + o_p(\rho_{\delta,N}). \)

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

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

Note that

\[
E[\hat{A}_{\text{sel},\text{int}}(\delta)] = E[(h + T_1^h + T_2^h)^2] - 2E[h^2 H^{-1}(T_1^H + T_2^H + T_3^H)]
\]

\[
= \sigma_u^2 H + \left(E[(h + T_1^h)^2] - \sigma_u^2 H\right) + 2E[(h + T_1^h)T_2^h] + E[(T_2^h)^2] - 2E[h^2 H^{-1}T_1^H]
\]

\[
- 2E[h^2 H^{-1}T_2^H] - 2E[h^2 H^{-1}T_3^H]
\]

\[
= \sigma_u^2 H + \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5 + \Delta_6.
\]

For the \( \Delta_1 \) term,

\[
\Delta_1 = E[(f^t P_{\delta} u)^2/N] - \sigma_u^2 H
\]

\[
= E \left[ \left( \sum_g \rho_g Z_g u_g 1(\hat{\mu}_g > \delta) \right)^2 \right] / N - \sigma_u^2 \sum_{g \in \mathcal{G}_{+\delta}} \mu_g^2 / N
\]

\[
= \sum_g E[\rho_g^2 (Z_g u_g)^2 1(\hat{\mu}_g > \delta)] / N + \left( \sum_g E[\rho_g Z_g u_g 1(\hat{\mu}_g > \delta)] \right)^2 / N
\]

\[
- \sum_g E[\rho_g Z_g^2 u_g 1(\hat{\mu}_g > \delta)]^2 - \sigma_u^2 \sum_{g \in \mathcal{G}_{+\delta}} \mu_g^2 / N
\]

\[
= \sigma_u^2 \sum_{g \in \mathcal{G}_{+\delta}} \mu_g^2 / N + \frac{\sigma_{uv}^2}{\sigma_v^2} \left( \sum_g \mu_g \phi_{g,\delta} \right)^2 / N
\]

\[
- \sigma_u^2 \sum_g \mu_g^2 \Phi_{g,\delta} / N + \frac{\sigma_{uv}^2}{\sigma_v^2} \sum_g \mu_g^2 \phi_{g,\delta} / N - \frac{\sigma_{uv}^2}{\sigma_v^2} \sum_g \mu_g^2 \phi_{g,\delta}^2 / N
\]

\[
= \frac{\sigma_{uv}^2}{\sigma_v^2} \left( \sum_g \mu_g \phi_{g,\delta} \right)^2 / N + O_p(G/N),
\]

52
where the last equality uses convergence results derived in Lemma A4.

For the $\Delta_2$ terms, we have

\[
\Delta_2 = 2E[(f'P_\delta u)(v'P_\delta u)/N] = 2E \left[ \sum_g \rho_g Z'_g u_g 1(\hat{\mu}_g \geq \delta) \sum_g v'_g Z'_g Z_g^{-1} Z'_g u_g 1(\hat{\mu}_g \geq \delta)/N \right]
\]

\[
= 2 \sum_g E \left[ \rho_g (Z'_g Z_g)^{-1}(Z'_g u_g)^2(Z'_g v_g)1(\hat{\mu}_g \geq \delta) \right] /N
\]

\[
+ 2 \sum_g E[\rho_g Z'_g u_g 1(\hat{\mu}_g \geq \delta)] \sum_g E[v'_g Z'_g (Z'_g Z_g)^{-1} Z'_g u_g 1(\hat{\mu}_g \geq \delta)] /N
\]

\[
- 2 \sum_g E[\rho_g Z'_g u_g 1(\hat{\mu}_g \geq \delta)] E[v'_g Z'_g (Z'_g Z_g)^{-1} Z'_g u_g 1(\hat{\mu}_g \geq \delta)] /N
\]

\[
= 2 \sigma^2_{uv} \sum_g \mu_g \phi_{g,\delta} (t_{g,\delta}^2 + 2) /N + 2\sigma_u(\sigma^2_u - \frac{\sigma^2_{uv}}{\sigma^2_v}) \sum_g \mu_g \phi_{g,\delta} /N
\]

\[
+ 2 \sigma^2_{uv} \sum_g \mu_g \phi_{g,\delta} \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta} \phi_{g,\delta}) /N - 2 \sigma^2_{uv} \sum_g \mu_g \phi_{g,\delta} (1 - \Phi_{g,\delta} + t_{g,\delta} \phi_{g,\delta}) /N
\]

\[
= 2 \sigma^2_{uv} \sum_g \mu_g \phi_{g,\delta} t_{g,\delta}^2 /N + 2\sigma_v \sigma^2_u \sum_g \mu_g \phi_{g,\delta} /N
\]

\[
+ 2 \sigma^2_{uv} \sum_g \mu_g \phi_{g,\delta} \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta} \phi_{g,\delta}) /N + 2 \sigma^2_{uv} \sum_g \mu_g \phi_{g,\delta} (\Phi_{g,\delta} - t_{g,\delta} \phi_{g,\delta}) /N
\]

\[
= 2 \sigma^2_{uv} \left( \sum_g \mu_g \phi_{g,\delta} \right) \left( \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta} \phi_{g,\delta}) \right) /N + O_p(G/N),
\]

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 $\Delta_3$ term, we have

$$
\Delta_3 = E[(v' P_g u)^2 / N] = E \left[ \left( \sum_g v'_g Z_g (Z'_g Z_g)^{-1} Z'_g u_g 1(\hat{\mu}_g > \delta) \right)^2 / N \right] = \sum_g E[(v'_g Z_g)^2 (Z'_g Z_g)^{-2} (Z'_g u_g)^2 1(\hat{\mu}_g > \delta)] / N + \left[ \sum_g E[v'_g Z_g (Z'_g Z_g)^{-1} Z'_g u_g 1(\hat{\mu}_g > \delta)] \right]^2 / N
$$

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

For the $\Delta_4$ term, notice that

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

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

For the $\Delta_5$ term,

$$
0 \leq \Delta_5 = 2E \left[ \left( \sum_g \rho_g Z'_g u_g \right)^2 \sum_g \rho_g^2 Z'_g Z_g 1(\hat{\mu}_g < \delta) \right] / (N^2 H) = 2 \sum_g E[\rho_g^4 (Z'_g u_g)^2 Z'_g Z_g 1(\hat{\mu}_g < \delta)] / (N^2 H)
$$

$$
+ 2 \sum_g E[\rho_g^2 (Z'_g u_g)^2] \sum_g E[\rho_g^2 Z'_g Z_g 1(\hat{\mu}_g < \delta)] / (N^2 H)
$$

$$
- 2 \sum_g E[\rho_g^2 (Z'_g u_g)^2] E[\rho_g^2 Z'_g Z_g 1(\hat{\mu}_g < \delta)] / (N^2 H)
$$

$$
= 2 \sum_g \rho_g^4 \left( \sigma_u^2 \Phi_{g, \delta} - \frac{\sigma^2_{u u}}{\sigma_v^2} \phi_{g, \delta} \right) / (N^2 H) + 2\sigma^2_u \sum_g \rho_g^2 \Phi_{g, \delta} (H + T_1^H) / (NH)
$$

$$
- 2\sigma^2_u \sum_g \rho_g^4 \Phi_{g, \delta} / (N^2 H) = O_p(G/N).
$$
Again, the last equality holds by convergence results derived in Lemma A4. Therefore, 
\[ \Delta_5 = O_p(G/N) . \]

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 \( \Delta_1 \) to \( \Delta_6 \), we get

\[
\sigma_{uv}^2 \left( \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta}) \right)^2 /N + 2\sigma_{uv}^2 \left( \sum_g \mu_g\phi_{g,\delta} \right) \left( \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta}) \right) /N \\
+ \sigma_{uv}^2 \left( \sum_g \mu_g\phi_{g,\delta} \right)^2 /N + O_p(G/N)
\]

\[
= \left( \sigma_{uv} \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta}) + \sigma_{uv} \sum_g \mu_g\phi_{g,\delta} \right)^2 /N + O_p(G/N).
\]

Notice that \( 0 \leq 1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta} \leq 1 \) and \( \sum_g \mu_g\phi_{g,\delta} = O_p(G) \) following convergence results derived in Lemma A4. Set \( S_{sel,int}(\delta)H^2 = \left( \sigma_{uv} \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta}) + \sigma_{uv} \sum_g \mu_g\phi_{g,\delta} \right)^2 /N \), we know that \( \rho_{\delta,N} = O_p(G^2/N) \). Further, since \( \sum_{g \in G_{+,s}} (\Phi_{g,\delta} + t_{g,\delta}\phi_{g,\delta}) = O_p(G^3/N^2) \) following convergence results derived in Lemma A4, we also know that \( S_{sel,int}(\delta)H^2 \geq \sigma_{uv}^2 G_{+,s}^2/N + o_p(G^2/N) = \sigma_{uv}^2 b^2 G^2/N + o_p(G^2/N) \) for some strictly positive \( b \) following Assumption 1. Therefore, any term of order \( o_p(G^2/N) \) is also of order \( o_p(\rho_{\delta,N}) \).

**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), \( \rho_{\delta,N} = o_p(1) \) as \( G^2/N \rightarrow 0 \).

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

\[
E[(T^h_1)^2] \leq \sqrt{E[(T^h_1)^2]} = \sigma^2_u \sum_g \mu_g^2 \Phi_{g,\delta}/N - \frac{\sigma^2_{uv}}{\sigma^2_v} \sum_g \mu_g^2 t_{g,\delta} \varphi_{g,\delta}/N
\]

\[
- \frac{\sigma^2_{uv}}{\sigma^2_v} \sum_g \mu_g^2 \varphi_{g,\delta}^2/N + \frac{\sigma^2_{uv}}{\sigma^2_v} \left[ \sum_g \mu_g \varphi_{g,\delta} \right]^2/N
\]

\[= O_p(G^2/N) = o_p(1). \]

Similarly \( T^h_1 = o_p(1) \) follows from the fact that \( \Delta_3 = O_p(G^2/N) = o_p(1). \)

For statement (4), note that \( T^H_1 = O_p(G/N) \) by following Assumption 1. \( T^H_2 = O_p(G/N) \) by Markov inequality and the fact that \( E[|T^H_2|] = \frac{1}{N} \sum_g \mu_g^2 \Phi_{g,\delta} = O_p(G/N) \). \( T^H_3 = O_p(1/\sqrt{N}) \) by the central limit theorem. Since each of \( G^2/N^2, N^{-1} \) and \( G/N/\sqrt{N} \) is \( o_p(G^2/N) \), \( (T^H_1 + T^H_2 + T^H_3)^2 = o_p(\rho_{\delta,N}). \)

For statement (5), note that \( T^h_1 + T^h_2 = O_p(G^2/N) \) and \( T^H_1 + T^H_2 + T^H_3 = O_p(G/N + 1/\sqrt{N}). \) Therefore, their product is \( o_p(G^2/N) = o_p(\rho_{\delta,N}). \)

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

\[
E \left[ (v'(I - P_\delta)f/N)^2 \right] = E \left[ \left( \sum_g \rho_g Z'_g v_g 1(\hat{\mu}_g < \delta) \right)^2 \right]/N^2
\]

\[= \sum_g \rho_g^2 E[(Z'_g v_g)^2 1(\hat{\mu}_g < \delta)]/N^2 + \left( E \left[ \sum_g \rho_g Z'_g v_g 1(\hat{\mu}_g < \delta) \right] \right)^2/N^2
\]

\[- \sum_g \left( E \left[ \rho_g Z'_g v_g 1(\hat{\mu}_g < \delta) \right] \right)^2/N^2
\]

\[= \sum_g \mu_g^2 \sigma^2_v (\Phi_{g,\delta} - t_{g,\delta} \varphi_{g,\delta})/N^2 + \left( \sum_g \mu_g \varphi_{g,\delta} \right)^2/N^2 - \sum_g (\mu_g \sigma_v \varphi_{g,\delta})^2/N^2
\]

\[= O_p(G^2/N^2), \]

by results in Lemma A4. Putting together, we have \( Z_H = o_p(\rho_{\delta,N}). \)

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 \( \mu^a_g = \rho_g \sqrt{(Z_g)'Z_g}, \) \( \hat{\mu}^a_g = ((Z_g)'Z_g)^{-1/2}(Z_g)'W_g, \)
\( \Phi^a_{g,\delta} = \Phi(\frac{\delta - \mu^a_g}{\sigma_v}), \) \( \phi^a_{g,\delta} = \phi(\frac{\delta - \mu^a_g}{\sigma_v}), \) \( t^a_g,\delta = \frac{\delta - \mu^a_g}{\sigma_v}, \) and define similar expressions for subsample b.

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

\[
N^a(\hat{\beta}^a(\delta) - \beta)^2 = \tilde{Q}^a(\delta) + \tilde{r}^a(\delta),
\]

\[
E[\tilde{Q}^a(\delta)|\tilde{Z}, X] = \sigma^2_a(H^a)^{-1} + S^a(\delta) + T^a(\delta),
\]

\[
(\tilde{r}^a(\delta) + T^a(\delta))/S^a(\delta) = o_p(1),
\]

with \( H^a = \sum_g \rho^2_g Z_g^a Z^a_g/N_a = \sum_g (\mu^a_g)^2/N_a \) and \( (H^a)^2 S^a(\delta) = \sigma^2_a \sigma_e^2 \sum_g (1 - \Phi^b_{g,\delta} + t^b_{g,\delta} \phi^b_{g,\delta})/N^a + \sigma^2_e \sum_g (\mu^a_g)^2 \Phi^a_{g,\delta}/N^a. \)

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

First, notice that \( \sqrt{N^a(\hat{\beta}^a(\delta) - \beta)} = (\hat{H}^a)^{-1} \hat{h}^a, \) where \( \hat{h}^a = \sum_g W_g^b Z_g^a (Z_g^b Z_g^b)^{-1} Z_g^a u_g^a (\hat{\mu}^b > \delta)/\sqrt{N^a} \) and \( \hat{H}^a = \sum_g W_g^b Z_g^a (Z_g^b Z_g^b)^{-1} Z_g^a W_g^b 1(\hat{\mu}^b > \delta)/N^a. \) Let \( \hat{h}^a = h^a + T^a_{1h} + T^a_{2h} \) and
\( \hat{H}^a = H^a + T^a_{1H} + T^a_{2H} + \hat{Z}_H, \) where

\[
h^a = \sum_g \rho_g Z_g^a u_g^a/\sqrt{N^a}; \quad H^a = \sum_g \rho^2_g Z_g^a Z^a_g/N^a; \quad T^a_{1h} = - \sum_g \rho_g 1(\hat{\mu}^b < \delta) Z_g^a u_g^a/\sqrt{N^a};
\]

\[
T^a_{2h} = \sum_g v_g^b Z_g^a (Z_g^b Z_g^b)^{-1} Z_g^a u_g^a 1(\hat{\mu}^b \geq \delta)/\sqrt{N^a}; \quad T^a_{1H} = - \sum_g \rho^2_g Z_g^a Z^a_g 1(\hat{\mu}^b < \delta)/N^a;
\]

\[
T^a_{2H} = \left( \sum_g \rho_g v_g^b Z_g^a (Z_g^b Z_g^b)^{-1} Z_g^a Z_g^a 1(\hat{\mu}^b \geq \delta) + \sum_g \rho_g Z_g^a v_g^a \right)/N^a;
\]

\[
Z_H = \sum_g v_g^b Z_g^a (Z_g^b Z_g^b)^{-1} Z_g^a Z_g^a u_g^a 1(\hat{\mu}^b \geq \delta)/N^a - \sum_g \rho_g Z_g^a v_g^a 1(\hat{\mu}^b < \delta)/N^a.
\]

Conforming to notations in Lemma A6 and let \( Z^4(\delta) = 0 \) hence \( \hat{A}^a(\delta) = (h^a + T^a_{1h} + T^a_{2h})^2 - 2(h^a)^2 (H^a)^{-1}(T^a_{1H} + T^a_{2H}). \) 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) $h^a = O_p(1), \ H^a = O_p(1), \ 2) \ \rho^a_{\delta,N} = S^a(\delta) = o_p(1), \ 3) \ T^a_{1h} + T^a_{2h} = o_p(1),$
4) $(T^a_{1H} + T^a_{2H})^2 = o_p(\rho^a_{\delta,N}), \ 5) \ (T^a_{1H} + T^a_{2H})(T^a_{1h} + T^a_{2h}) = o_p(\rho^a_{\delta,N}),$
6) $Z^a_H = o_p(\rho^a_{\delta,N}), \ 7) \ E[\hat{A}^a(\delta)] = \sigma^2_u H^a + (H^a)^2 S^a(\delta) + o_p(\rho^a_{\delta,N}).$

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

We have

$$E[\hat{A}^a(\delta)] = \sigma^2_u H^a + (E[(h^a + T^a_{1h})^2] - \sigma^2_u H^a) + 2E[(h^a + T^a_{1h})T^a_{2h}] + E[(T^a_{2h})^2] - 2E[(h^a)^2 T^a_{1H}/H^a] - 2E[(h^a)^2 T^a_{2H}/H^a]$$

$$= \sigma^2_u H^a + \Delta^a_1 + \Delta^a_2 + \Delta^a_3 + \Delta^a_4 + \Delta^a_5.$$

For the $\Delta^a_1$ term,

$$\Delta^a_1 = E \left[ \left( \sum_g \rho_g 1(\mu^b_g \geq \delta) Z^a_g u^a_g \right)^2 \right] / N^a - \sigma^2_u H^a$$

$$= \sum_g \rho_g^2 E[1(\mu^b_g \geq \delta)(Z^a_g u^a_g)^2]/N^a - \sigma^2_u H^a = \sigma^2_u \sum_g (\mu^a_g)^2 (1 - \Phi^b_{g,\delta})/N^a - \sigma^2_u H^a$$

$$= -\sigma^2_u \sum_g (\mu^a_g)^2 \Phi^b_{g,\delta}/N^a.$$

For the $\Delta^a_2$ term,

$$\Delta^a_2 = -2\rho_g E \left[ \sum_g (Z^a_g u^a_g 1(\mu^b_g \geq \delta)) \sum_g (v^b_g)^t Z^b_g ((Z^b_g)^t Z^b_g)^{-1} (Z^b_g)^t u^b_g 1(\mu^b_g \geq \delta)/N \right]$$

$$= -2 \sum_g \rho_g E[((Z^a_g u^a_g)^2] E [(v^b_g)^t Z^b_g ((Z^b_g)^t Z^b_g)^{-1} (Z^b_g)^t u^b_g 1(\mu^b_g \geq \delta)/N]$$

$$= -2\sigma^2_u \sum_g \rho_g (Z^a_g)^t Z^a_g \Phi^b_{g,\delta} / (Z^b_g)^t Z^b_g$$

$$= -2\sigma^2_u \sum_g \mu^a_g \Phi^b_{g,\delta} \sqrt{((Z^a_g)^t Z^a_g) / ((Z^b_g)^t Z^b_g))}$$

58
For the $\Delta_3^a$ term,

$$\Delta_3^a = E \left[ \left( \sum_g (\mu_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1} (Z_g^a)' u_g^a 1(\mu_g^b \geq \delta) \right)^2 \right] / N^a$$

$$= \sum_g E \left[ ((\mu_g^b)' Z_g^b ((Z_g^b)' Z_g^b)^{-1} (Z_g^a)' u_g^a)^2 1(\mu_g^b \geq \delta) \right] / N^a$$

$$= \sum_g E \left[ ((\mu_g^b)' Z_g^b)^2 1(\mu_g^b \geq \delta) \right] / ((Z_g^b)' Z_g^b)^2 E [(Z_g^a)' u_g^a]^2] / N^a$$

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

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

$$= \sigma_v^2 \sigma_u^2 \sum_g (1 - \Phi_{g,\delta} + t_{g,\delta} \phi_{g,\delta}) / N^a + o_p(G/N).$$

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 $\eta_g$ is the residual of a linear projection of $\bar{Z}_g$ onto $X_g$. Therefore $Z_g' Z_g = \eta_g' \eta_g - \eta_g' P_{X_g} \eta_g$. Under Assumption 1 we have $E[\eta_g^2] = k_g$ and $E[(\eta_g^2 - k_g)^2] \leq \Delta_\eta < \infty$ 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 $C^2 = \Delta_\eta \epsilon$, then

$$P \left( \sum_i (\eta_{ig}^a - k_g) \right) \geq C(N/G)^{1/2} \leq \frac{E \left( \sum_i (\eta_{ig}^a - k_g)^2 \right)}{C^2 N/G} \leq \frac{2 \sum_i E \left( \eta_{ig}^a - k_g \right)^2}{C^2 N/G} \leq \frac{1}{C^2} \Delta_\eta \epsilon = \epsilon$$

This implies that $Z_g^a Z_g^a / \eta_g^a - k_g = O_p(1/\sqrt{N/G})$ 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 $\Delta_4^a$ term,

$$\Delta_4^a = 2E \left[ \sum_g \rho_g (Z_g^a)' u_g^a \right]^2 \sum_g \rho_g^2 (Z_g^a)' Z_g^a 1(\mu_g^b < \delta) \right] / ((N^a)^2 H^a)$$

$$= 2E \left[ \left( \sum_g \rho_g (Z_g^a)' u_g^a \right) \sum_g (\mu_g^a)^2 \right] 1(\mu_g^b < \delta) \right] / ((N^a)^2 H^a)$$

$$= 2 \sum_g \rho_g^2 E \left[ (Z_g^a)' u_g^a \right] \sum_g (\mu_g^a)^2 \Phi_{g,\delta} / ((N^a)^2 H^a) = 2 \sigma_u^2 \sum_g (\mu_g^a)^2 \Phi_{g,\delta} / N^a.$$
For the $\Delta^a_5$ term,

$$\Delta^a_5 = -2E \left[ \left( \sum_g \rho_g(Z^a_g)'u^a_g \right)^2 \sum_g \rho_g(v^b_g)'Z^b_g((Z^b_g)'Z^b_g)^{-1}(Z^a_g)'Z^a_g 1(\hat{\mu}^b_g \geq \delta) \right] / ((N^a)^2H^a)$$

$$-2E \left[ \left( \sum_g \rho_g(Z^a_g)'u^a_g \right)^2 \sum_g \rho_g(Z^a_g)'t^a_g \right] / ((N^a)^2H^a)$$

$$= -2E \left[ \left( \sum_g \rho_g(Z^a_g)'u^a_g \right)^2 \sum_g E \left[ \rho_g(v^b_g)'Z^b_g((Z^b_g)'Z^b_g)^{-1}(Z^a_g)'Z^a_g 1(\hat{\mu}^b_g \geq \delta) \right] / ((N^a)^2H^a)$$

$$= -2\sigma^2_u \sum_g \rho^b_g \phi^b_{g,\delta} \sqrt{((Z^a_g)'Z^a_g)/((Z^b_g)'Z^b_g)}/N^a.$$ 

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

Collecting the leading terms from $\Delta^a_1$ to $\Delta^a_5$ we get

$$\sigma^2_v \sigma^2_u \sum_g (1 - \Phi^b_{g,\delta} + t^b_{g,\delta} \phi^b_{g,\delta}) / N^a + \sigma^2_u \sum_g (\mu^a_g)^2 \Phi^b_{g,\delta} / N^a,$$

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 $S^a(\delta) = O_p(G/N)$. In addition, $\sum_{g \in G_{\deltaN}} (\Phi^a_{g,\delta} - t^a_{g,\delta} \phi^a_{g,\delta}) / N = O_p(G^3/N^3)$. Therefore, $S^a(\delta) \geq \sigma^2_v \sigma^2_u bG/N + o_p(G/N)$ and any terms of order $o_p(G/N)$ is also of order $o_p(\rho^a_{\delta,n})$.

**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 2.

For statement (2), $S^a(\delta) = o_p(1)$ as $G/N \rightarrow 0$.

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

To prove (4), notice that $T^a_{1H} = O_p(G/N)$ by the Markov inequality and the fact that $E[|T^a_{1h}|] = E \left[ \sum_g \rho^2_g Z^a_g'Z^a_g 1(\hat{\mu}^b_g < \delta) \right] / N^a = \sum_g (\mu^a_g)^2 \Phi^b_{g,\delta} / N^a$.

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

\[
E\left[\left(\sum_g \rho_g (v_g^b)^\prime Z_g^b ((Z_g^b)^\prime Z_g^b)^{-1} (Z_g^a)^\prime Z_g^a 1(\hat{\mu}_g^b > \delta)\right)^2\right]/(N^a)^2
\]

\[
= \sigma_v^2 \sum_g (\hat{\mu}_g^b)^2 (1 - \Phi^b_{g,\delta} + \sigma^b_{g,\delta}) + \left(\sum_g \mu_g^b \sigma_{g,\delta}\right)^2 - \sum_g (\hat{\mu}_g^b)^2 (\sigma_{g,\delta})^2 \left((Z_g^a)^\prime Z_g^a / (Z_g^b)^\prime Z_g^b\right) / (N^a)^2
\]

\[
= O_p(G^2/N^2).
\]

Since each of $1/N$, $G^2/N^2$, and $G/N/\sqrt{N}$ is of order $o_p(G/N) = o_p(S^a(\delta))$, hence $(T_{1H}^a + T_{2H}^a)^2 = o_p(\rho_{\delta,N}^a)$.

For statement (5), note that by statements (3) and (4) $T_{1h}^a + T_{2h}^a = o_p(1)$, and $T_{1H}^a + T_{2H}^a = O_p(G/N)$. Therefore $(T_{1h}^a + T_{2h}^a)(T_{1H}^a + T_{2H}^a) = o_p(G/N) = o_p(\rho_{\delta,N}^a)$.

Lastly, for statement (6), notice that the first term of $Z_H^a$ is of order $O_p(\sqrt{G/N}) = o_p(\rho_{\delta,N}^a)$ by the Markov inequality, the Cauchy-Schwarz inequality and the facts that

\[
E[\{Z^a_{1H}\}^2] = E\left[\left(\sum_g (v_g^b)^\prime Z_g^b ((Z_g^b)^\prime Z_g^b)^{-1} (Z_g^a)^\prime v_g^a 1(\hat{\mu}_g^b > \delta)\right)^2\right]/(N^a)^2
\]

\[
= \sum_g E \left[((v_g^b)^\prime Z_g^b ((Z_g^b)^\prime Z_g^b)^{-1})^2 1(\hat{\mu}_g^b > \delta)\right] E \left[(Z_g^a)^\prime v_g^a\right] / (N^a)^2
\]

\[
= \sigma_v^4 \sum_g (1 - \Phi^b_{g,\delta} + \sigma^b_{g,\delta}) (Z_g^a)^\prime Z_g^a / ((Z_g^b)^\prime Z_g^b) / (N^a)^2 = O_p\left(\frac{G}{N^2}\right),
\]

\[
E[\{Z^a_{2H}\}^2] = E\left[\left(\sum_g \rho_g (Z_g^a)^\prime v_g^a 1(\hat{\mu}_g^b < \delta)\right)^2\right]/(N^a)^2 = \sigma_v^2 \sum_{g \in G} (\mu_g^a)^2 \Phi_{g,\delta} / (N^a)^2
\]

Putting together, we have $Z_H^a = o_p(\rho_{\delta,N}^a)$.

Following the three steps, the lemma is hence proven.

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

Note that by Lemma A7, we know that

\[
\sqrt{N^a} (\hat{\rho}_{\text{asset, int}}^a (\delta) - \beta) = (H_H^a)^{-1} \hat{h}_H^a
\]

\[
= (H^a)^{-1} (h^a + T_{1h}^a + T_{2h}^a - (T_{1H}^a + T_{2H}^a)(H^a)^{-1} h^a) + o_p(G/N),
\]

61
and a similar result hold for $\hat{\beta}_{ssel,int}^b(\delta)$.

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 $N^b$. Further, since $\hat{H}_g^a$, $\hat{h}_g^a$, $\hat{H}_g^b$, $\hat{h}_g^b$ are all $O_p(1)$, we know that

$$\sqrt{N}(\hat{\beta}_{ssel,int}(\delta) - \beta) = \frac{1}{2} \left( \sqrt{N/N^a} \sqrt{N^a(\hat{\beta}_{ssel,int}^a(\delta) - \beta)} + \sqrt{N/N^b} \sqrt{N^b(\hat{\beta}_{ssel,int}^b(\delta) - \beta)} \right)$$

$$= (h^a + T_{1h}^a + T_{2h}^a - (T_{1h}^a + T_{2h}^a)(H^a)^{-1}h^a)(h^b + T_{1h}^b + T_{2h}^b - (T_{1h}^b + T_{2h}^b)(H^b)^{-1}h^b)$$

$$+ (h^b + T_{1h}^b + T_{2h}^b - (T_{1h}^b + T_{2h}^b)(H^b)^{-1}h^b) / H^a / \sqrt{2}$$

$$+ (h^b + T_{1h}^b + T_{2h}^b - (T_{1h}^b + T_{2h}^b)(H^b)^{-1}h^b) / H^b / \sqrt{2}$$

$$+ o_p(G/N).$$

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

$$= (h^a + T_{1h}^a + T_{2h}^a)(h^b + T_{1h}^b + T_{2h}^b) - (h^a + T_{1h}^a + T_{2h}^a)(T_{1h}^b + T_{2h}^b)(H^b)^{-1}h^b$$

$$- (T_{1h}^a + T_{2h}^a)(H^a)^{-1}h^a(h^b + T_{1h}^b + T_{2h}^b) + (T_{1h}^a + T_{2h}^a)(H^a)^{-1}h^a(T_{1h}^b + T_{2h}^b)(H^b)^{-1}h^b$$

$$= (h^a + T_{1h}^a + T_{2h}^a)(h^b + T_{1h}^b + T_{2h}^b) - h^a(T_{1h}^a + T_{2h}^a)(H^a)^{-1}h^b - (T_{1h}^a + T_{2h}^a)(H^a)^{-1}h^a h^b$$

$$+ o_p(G/N),$$

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

$$\hat{Q}_{ab}^a(\delta) H^a H^b = (h^a + T_{1h}^a + T_{2h}^a)(h^b + T_{1h}^b + T_{2h}^b)$$

$$- h^a(T_{1h}^a + T_{2h}^a)(H^b)^{-1}h^b - (T_{1h}^a + T_{2h}^a)(H^a)^{-1}h^a h^b.$$  

Then we know

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

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 $E[h^a(T_{1h}^b + T_{2h}^b)h^b] = 0$, we know that

$$E[\hat{Q}_{ab}^a(\delta) H^a H^b = E[(h^a + T_{1h}^a + T_{2h}^a)(h^b + T_{1h}^b + T_{2h}^b)] = E[(T_{1h}^a + T_{2h}^a)(T_{1h}^b + T_{2h}^b)]$$

62
as \( E[h^a h^b] = E[h^a T_{1h}^b] = E[h^a T_{2h}^b] = E[T_{1h}^a h^b] = E[T_{2h}^a h^b] = 0 \). For the last expectation, we know that
\[
E[(T_{1h}^a + T_{2h}^a)(T_{1h}^b + T_{2h}^b)] = \sum_g E[(v_g^a)' P_{Z_g} u_g^a 1(\hat{\mu}_g^a \geq \delta)] E[(v_g^b)' P_{Z_g} u_g^b 1(\hat{\mu}_g^b \geq \delta)] / \sqrt{N_a N_b}
+ \sum_g \rho_g^2 E[1(\hat{\mu}_g^a < \delta)(Z_g^a)' u_g^a] E[1(\hat{\mu}_g^b < \delta)(Z_g^b)' u_g^b] / \sqrt{N_a N_b}
- \sum_g \rho_g E[(v_g^a)' P_{Z_g} u_g^a 1(\hat{\mu}_g^a \geq \delta)] E[1(\hat{\mu}_g^a < \delta)(Z_g^b)' u_g^b] / \sqrt{N_a N_b}
- \sum_g \rho_g E[(v_g^b)' P_{Z_g} u_g^b 1(\hat{\mu}_g^b \geq \delta)] E[1(\hat{\mu}_g^a < \delta)(Z_g^a)' u_g^a] / \sqrt{N_a N_b}
\]
\[
= \sum_g \sigma_{uv}^2 (1 - \Phi_{g, \delta}^a + t_{g, \delta}^a \phi_{g, \delta}^a) (1 - \Phi_{g, \delta}^b + t_{g, \delta}^b \phi_{g, \delta}^b) / (N/2) + \sum_g \sigma_{uv}^2 \sigma_v^2 \Sigma_g \mu_g^a \mu_g^b \phi_{g, \delta}^a \phi_{g, \delta}^b / (N/2)
+ \sigma_{uv}^2 \sigma_v^2 \sum_g \mu_g^b (1 - \Phi_{g, \delta}^a + t_{g, \delta}^a \phi_{g, \delta}^a) \phi_{g, \delta}^b + \mu_g^a (1 - \Phi_{g, \delta}^b + t_{g, \delta}^b \phi_{g, \delta}^b) \phi_{g, \delta}^a) / (N/2) + o_p(G/N)
= 2 \sigma_{uv}^2 \sum_g (1 - \Phi_{g, \delta}^a + \delta \phi_{g, \delta}^a) (1 - \Phi_{g, \delta}^b + \delta \phi_{g, \delta}^b) / N + o_p(G/N).
\]

Further, since \( |\Phi(\frac{\delta - \mu_g}{\sigma_v}/\sqrt{2}) - \Phi(\frac{\delta - \mu_g}{\sigma_v}/\sqrt{2})| \leq |\mu_g/\sqrt{2} - \mu_g^a| \phi(\frac{\delta - \mu_g^a}{\sigma_v}) \) for a \( \mu_g^a \) between \( \mu_g/\sqrt{2} \) and \( \mu_g \), by Cauchy-Schwartz inequality we have the following results:
\[
E \left[ \left( \frac{\delta - \mu_g}{\sigma_v} \right) - \Phi(\frac{\delta - \mu_g}{\sigma_v}) \right] \leq \sqrt{E \left[ \left( \frac{\mu_g}{\sqrt{2}} - \mu_g^a \right)^2 \right] \sqrt{E \left[ \phi^2(\frac{\delta - \mu_g^a}{\sigma_v}) \right]}}
\leq \rho_g^2 n_g \cdot \sqrt{E \left[ \left( \frac{\sqrt{Z_g^a Z_g/n_g} - \sqrt{(Z_g^a)' Z_g/a_g}}{\sqrt{(Z_g^a)' Z_g/a_g}} \right)^2 \right] \sqrt{E \left[ \phi^2(\frac{\delta - \mu_g^a}{\sigma_v}) \right]}}
\leq \sqrt{E \left[ \left( \frac{\sqrt{Z_g^a Z_g/n_g} - \sqrt{(Z_g^a)' Z_g/a_g}}{\sqrt{(Z_g^a)' Z_g/a_g}} \right)^2 \right]} \leq (G/N)^{-1/4}
\]
and
\[
\left| \frac{\delta - \mu_g}{\sigma_v} \phi(\frac{\delta - \mu_g}{\sigma_v}) - \frac{\delta - \mu_g^a}{\sigma_v} \phi(\frac{\delta - \mu_g^a}{\sigma_v}) \right| \leq |\mu_g/\sqrt{2} - \mu_g^a| \left| \frac{\delta - \mu_g^a}{\sigma_v} \phi(\frac{\delta - \mu_g^a}{\sigma_v}) \right| \leq (G/N)^{-1/4}
\]
and
\[
\left| \frac{\mu_g}{\sqrt{2}} \phi(\frac{\delta - \mu_g}{\sigma_v}) - \mu_g^a \phi(\frac{\delta - \mu_g^a}{\sigma_v}) \right| \leq |\mu_g/\sqrt{2} - \mu_g^a| \left| \frac{\delta - \mu_g^a}{\sigma_v} \phi(\frac{\delta - \mu_g^a}{\sigma_v}) \right| + |\mu_g/\sqrt{2} - \mu_g^a| \left| \frac{\delta - \mu_g^a}{\sigma_v} \phi(\frac{\delta - \mu_g^a}{\sigma_v}) \right| \leq (G/N)^{-1/4}
\]
Therefore, we know that

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

Similarly, we could also simplify the result in Lemma A7 and get that

\[ E[\hat{Q}^a(\delta)|\bar{Z}, X] + E[\hat{Q}^b(\delta)|\bar{Z}, X] \]

\[ = \sigma_u^2(1/H^a + 1/H^b)/2 \]

\[ + \sigma_v^2 \sum_g \left( 1 - \Phi \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) + \frac{\delta}{\sigma_v} \phi \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) \right) \left( 1/(H^a)^2 + 1/(H^b)^2 \right)/N \]

\[ + \sigma_v^2 \sum_g \mu_g^2 \Phi \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) \left( 1/(H^a)^2 + 1/(H^b)^2 \right)/N. \]

Moreover, since \( \frac{1}{N^2} \sum_g (Z_g^a)^2 Z_g^a = \frac{1}{N^2} \sum_g n_g^a ((Z_g^a)^2/n_g^a - k_g) + \frac{1}{N^2} \sum_g n_g^a k_g = k + o_p(G/N) = k + o_p(G/N) \) and similarly \( \sum_g Z_g^a Z_g/N = k + o_p(G/N) \). Therefore \( |H^a - H| \leq o_p(G/N) \) and we could further simplify the results and obtain the formulation of \( S_{ssel, int}(\delta) \) stated in the theorem.

\[ \square \]

**Proof of Corollary 1**

Recall the definition of \( A_{ssel, int}(\delta), B_{ssel, int}(\delta), \) and \( C_{ssel, int}(\delta) \) 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 \( C_1(\delta) = 2\sigma_{uv}^2 \sum_g \left( 1 - \Phi \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) \right) \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right)^2 \) and \( C_2(\delta) = 4\sigma_v^2 \left( 1 - \Phi \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) \right) \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right)^2 \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) \right)^2 \). It is clear that both \( C_1(\delta) \) and \( C_2(\delta) \) are non-negative as well.

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

and \( \nabla_\delta C_1(\delta) = -2\sigma_{uv}^2 \sum_g \left( 1 - \Phi \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) \right) \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right)^2 \phi \left( \frac{\delta - \mu_g/\sqrt{2}}{\sigma_v} \right) \leq 0. \) 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, use superscript to denote groups where summation is taken from. For example, \( L^0(\delta) \equiv A^0(\delta) + B^0(\delta) + C_1^0(\delta) + C_2^0(\delta) \) with \( A^0(\delta) = 2\sigma_u^2 \sum_{g \in G_0} \left( 1 - \Phi \left( \frac{\delta}{\sigma_v} \right) - \frac{\delta}{\sigma_v} \phi \left( \frac{\delta}{\sigma_v} \right) \right)^2 \), \( C_1^0(\delta) = 2\sigma^2_{u\sigma^2} \sum_{g \in G_0} \left( 1 - \Phi \left( \frac{\delta}{\sigma_v} \right) - \frac{\delta}{\sigma_v} \phi \left( \frac{\delta}{\sigma_v} \right) \right)^2 \), and \( B^0(\delta) = C_2^0(\delta) = 0 \). To prove the Corollary, we show in sequence the following results:

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

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

3) \( L(\delta) \geq 2b\sigma_0^2 \sigma_v^2 \left( 1 + \rho_{uw}^2 \right) \frac{G}{N} + o_p\left( \frac{G}{N} \right) \) for \( \forall \delta \geq 0 \).

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

\[
L^0(\delta_N) = A^0(\delta_N) + C_1^0(\delta_N) \\
= 2\sigma_u^2 \sigma_v^2 \frac{G_0}{N} \left( 1 - \Phi \left( \frac{\delta_N}{\sigma_v} \right) - \frac{\delta_N}{\sigma_v} \phi \left( \frac{\delta_N}{\sigma_v} \right) \right)^2 + 2\sigma^2_{u\sigma^2} \frac{G_0}{N} \left( 1 - \Phi \left( \frac{\delta_N}{\sigma_v} \right) - \frac{\delta_N}{\sigma_v} \phi \left( \frac{\delta_N}{\sigma_v} \right) \right)^2 \\
\leq 2\sigma_u^2 \sigma_v^2 \frac{G_0}{N} \left( 1 + \frac{\delta_N}{\sigma_v} \phi \left( \frac{\delta_N}{\sigma_v} \right) \right)^2 + 2\sigma^2_{u\sigma^2} \frac{G_0}{N} \left( 1 + \frac{\delta_N}{\sigma_v} \phi \left( \frac{\delta_N}{\sigma_v} \right) \right)^2 \\
\leq \sigma_u^2 \sigma_v^2 \frac{G_0}{N} \left( \frac{\delta_N}{\sigma_v} \right)^3 + 1 \left( \frac{\delta_N}{\sigma_v} \right)^3 + \frac{G_0}{N} \left( \frac{\delta_N}{\sigma_v} \right)^3 + o(G/N) \\
\]

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

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

and

\[
A^{+,w}(\delta) + C_1^{+,w}(\delta) \leq \frac{\sigma^2_u \sigma_v^2}{N} \sum_{g \in G^{+,w}} \left( \frac{\delta - \mu_g / \sqrt{2}}{\sigma_v} \right)^3 + \frac{\sigma^2_{u\sigma^2}}{N} \sum_{g \in G^{+,w}} \left( \frac{\delta - \mu_g / \sqrt{2}}{\sigma_v} \right)^2 \\
= \frac{G^{+,w}}{G} O_p(G/N), 
\]

65
Similarly,
\[
C_{2}^{+,w}(\delta) \leq \sum_{g \in \mathcal{G}_{+,w}} \frac{2\sigma_{uw}^{2} \mu_{g}}{N} \frac{\sqrt{2}}{\sigma_{v}} \phi \left( \frac{\delta - \mu_{g}/\sqrt{2}}{\sigma_{v}} \right) + \sum_{g \in \mathcal{G}_{+,w}} \frac{\sigma_{uv}^{2}}{N\sigma_{v}^{2}} \mu_{g}^{2} \phi^{2} \left( \frac{\delta - \mu_{g}/\sqrt{2}}{\sigma_{v}} \right)
\]
\[
= \frac{G_{+,w}}{G} O_{p}(G/N).
\]

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 \(\mathcal{U} \equiv 2\sigma_{u}^{2}\sigma_{v}^{2}(1 + \rho_{uv}^{2})\). Let \(\delta_{g}^{*}\) satisfy that \(\sigma_{u}^{2}\mu_{g}^{2} \Phi \left( \frac{\delta_{g}^{*} - \mu_{g}/\sqrt{2}}{\sigma_{v}} \right) = \mathcal{U}\). Then it is easy to derive that \(\Phi \left( \frac{\delta_{g}^{*} - \mu_{g}/\sqrt{2}}{\sigma_{v}} \right) = \frac{\mathcal{U}}{\sigma_{u}^{2}\mu_{g}^{2}}\), or that \(\delta_{g}^{*} = \frac{\mu_{g}/\sqrt{2} + \sigma_{v} \Phi^{-1} \left( \frac{\mathcal{U}}{\sigma_{u}^{2}\mu_{g}^{2}} \right)}{}\).

For any \(\delta > \delta_{g}^{*}\), since \(B(\delta)\) increases in \(\delta\), we know that \(B^{+,s}(\delta) \geq \frac{G_{+,s}}{G} \cdot \frac{\mathcal{U}}{N}\). For any \(0 \leq \delta \leq \delta_{g}^{*}\), notice that

\[
\left| A^{+,s}(\delta) + C^{+,s}(\delta) - \frac{G_{+,s}}{G} \cdot \frac{\mathcal{U}}{N} \right| \leq 2 \frac{\sigma_{u}^{2}\sigma_{v}^{2}}{N} \sum_{g \in \mathcal{G}_{+,s}} |A_{g} - 1| + 2 \frac{\sigma_{uv}^{2}}{N} \sum_{g \in \mathcal{G}_{+,s}} |A_{g}^{2} - 1|
\]
\[
\leq 2 \frac{\sigma_{u}^{2}\sigma_{v}^{2}}{N} \sum_{g \in \mathcal{G}_{+,s}} (1 - A_{g}) + 4 \frac{\sigma_{uv}^{2}}{N} \sum_{g \in \mathcal{G}_{+,s}} (1 - A_{g}) \leq 6 \frac{\sigma_{u}^{2}\sigma_{v}^{2}}{N} \sum_{g \in \mathcal{G}_{+,s}} (1 - A_{g})
\]
\[
\leq 6 \frac{\sigma_{u}^{2}\sigma_{v}^{2}}{N} \sum_{g \in \mathcal{G}_{+,s}} (1 - A_{g}) = 6 \frac{\sigma_{u}^{2}\sigma_{v}^{2}}{N} \sum_{g \in \mathcal{G}_{+,s}} \left( \Phi(x_{g}^{*}) - x_{g}^{*} \phi(x_{g}^{*}) \right)
\]
\[
= \mathcal{O}_{p} \left( \frac{1}{N} \sum_{g \in \mathcal{G}_{+,s}} E \left[ \Phi(x_{g}^{*}) \right] \right) + \mathcal{O}_{p} \left( \frac{1}{N} \sum_{g \in \mathcal{G}_{+,s}} E \left[ x_{g}^{*} \phi(x_{g}^{*}) \right] \right) = o_{p}(G/N),
\]

where \(A_{g} = 1 - \Phi \left( \frac{\delta_{g} - \mu_{g}/\sqrt{2}}{\sigma_{v}} \right) + \frac{\delta_{g} - \mu_{g}/\sqrt{2}}{\sigma_{v}} \Phi \left( \frac{\delta_{g} - \mu_{g}/\sqrt{2}}{\sigma_{v}} \right)\) which increases with \(\delta\) and let \(x_{g}^{*} = (\delta_{g}^{*} - \mu_{g}/\sqrt{2})/\sigma_{v} = \Phi^{-1} \left( \frac{\mathcal{U}}{\sigma_{u}^{2}\mu_{g}^{2}} \right)\) and \(A_{g}^{*} = 1 - \Phi(x_{g}^{*}) + x_{g}^{*} \phi(x_{g}^{*}) \leq 1\).

The last equality follows as

\[
E \left[ \Phi(x_{g}^{*}) \right] = E \left[ \Phi(x_{g}^{*}) \right] 1 \left( \mu_{g} > \rho_{g} \sqrt{k_{g} n_{g}/2} \right) + P(\mu_{g} \leq \rho_{g} \sqrt{k_{g} n_{g}/2})
\]
\[
= E \left[ \frac{\mathcal{U}}{\sigma_{u}^{2}\mu_{g}}^2 1 \left( \mu_{g} > \rho_{g} \sqrt{k_{g} n_{g}/2} \right) \right] + P(\mu_{g} \leq \rho_{g} \sqrt{k_{g} n_{g}/2})
\]
\[
\leq \frac{2\sigma_{u}^{2}(1 + \rho_{uv}^{2})}{\rho_{g}^{2}k_{g} n_{g}/2} + P \left( Z_{g}^{2} n_{g} \leq k_{g}/2 \right) \lesssim G/N,
\]

66
and
\[
E \left[ x_g^* \phi(x_g^*) \right] \leq E \left[ x_g^* \phi(x_g^*) \left( \mu_g > \rho_g \sqrt{k_n g/2} \right) \right] + P \left( \mu_g \leq \rho_g \sqrt{k_n g/2} \right) \\
\leq E \left[ x_g^* \phi(x_g^*) \right] + P \left( Z' Z/n_g \leq k_g/2 \right) \\
\leq E \left[ 1/|x_g^*|/\sqrt{2\pi} \right] + P \left( Z' Z/n_g \leq k_g/2 \right) \\
\leq \left| \Phi^{-1} \left( 2\sigma^2(1 + \rho_{uv}^2) \right) \right| + P \left( Z' Z/n_g \leq k_g/2 \right) = o(1)
\]
where \( x_g^* = \Phi^{-1} \left( 2\sigma^2(1 + \rho_{uv}^2) \right) \) and last convergence result holds uniformly over \( g \).

Therefore, we know that

\[
A^{+,s}(\delta) + C_1^{+,s}(\delta) = \mathcal{U} G^{+,s}_G + o_p(G/N).
\]

Noticing that \( A(\delta), B(\delta), C_1(\delta), \) and \( C_2(\delta) \) are positive with all groups and since \( G^{+,s}_G \to b \) under Assumption 1, we therefore know that \( L(\delta) \geq 2b \sigma^2 \sigma^2_v (1 + \rho_{uv}^2) + o_p(G/N) \) for \( \forall \delta \geq 0 \).

**Proof of Theorem 3**

Since \( G^{+,w} = 0 \), conditional on sample path, there exists an ordering of \( \mu_g \) denoted by

\[
\mu(1) \geq \mu(2) \cdots \geq \mu(G^{+,s}) > \mu(G^{+,s} + 1) = \cdots = \mu(G) = 0.
\]

Define

\[
\mathcal{R}(K) \equiv \frac{\sigma^2_u}{N} \sum_{g=K+1}^{G} \mu^2_{(g)} + 2\sigma^2_u \sigma^2_v (1 + \rho_{uv}^2) \frac{K}{N}.
\]

It is easy to show that there is a link between \( \mathcal{R}(K) \) and \( L(\delta) \) defined in the proof of Coroly 1 such that

\[
\frac{\min_K \mathcal{R}(K)}{\inf_{\delta} L(\delta)} \xrightarrow{p} 1.
\]

This is because on one hand, as \( G^{+,w} = 0 \), \( \mathcal{R}(K) \) decreases in \( K \) for \( K \leq G^{+,s} \) while on the other, since \( \min_{g \in G^{+,s}} \mu_g^2 \geq \frac{N}{G} \) by Lemma A5, \( \mathcal{R}(K) \) increases in \( K \) for \( K > G^{+,s} \) with large enough \( N \). Hence \( \min_K \mathcal{R}(K) = \mathcal{R}(G_1) = 2b \sigma^2 u \sigma^2_v (1 + \rho_{uv}^2) \cdot \frac{G}{N} + o(\frac{G}{N}) \).

Also define its estimator

\[
\hat{\mathcal{R}}(K) = \frac{\hat{\sigma}^2_u}{N} \sum_{g=K+1}^{G} \hat{\mu}^2_{(g)} + 2\hat{\sigma}^2_u \hat{\sigma}^2_v (1 + \hat{\rho}_{uv}^2) \frac{K}{N} \sqrt{\frac{N}{G} \log G}
\]

67
where $\hat{\mu}_1 \geq \hat{\mu}_2 \cdots \geq \hat{\mu}_G$ with $\hat{\mu}_g = \hat{\rho}_g \sqrt{Z'_g Z_g} = (Z'_g Z_g)^{-1/2} Z'_g v_g$ and $\hat{\sigma}_g^2$, $\hat{\sigma}_u^2$, $\hat{\rho}_{uv}$ are consistent estimators for $\sigma_g^2$ and $\sigma_u^2$ and $\rho_{uv}$ respectively using the whole sample. For instance, we can construct these using residuals of the first stage and second stage regression based on the pooled 2SLS estimator.

Let $\hat{K} = \arg\min_K \hat{R}(K)$, we can establish the following statement

$$\frac{R(\hat{K})}{\min_K R(K)} \xrightarrow{p} 1.$$ 

To prove this statement, we just need to prove that as $N \to +\infty$, with probability goes to 1,

i) $\frac{\min_{g \in \mathcal{G}_{+,s}} \mu_g^2}{\sqrt{\frac{N}{2} \log G}} \to +\infty$

ii) $\frac{\max_{g \in \mathcal{G}_{+,s}} \mu_g^2}{\sqrt{\frac{N}{2} \log G}} \to 0$

This is because statements (i) and (ii) imply that $\hat{R}(K)$ decreases for $K \leq G_1$ while $\hat{R}(K)$ is increasing for $K > G_1$ and thus $\hat{K} = \arg\min_K \hat{R}(K) = G_1$ with probability goes to 1.

The rest of the proof are for statements (i) and (ii). To prove i), note

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

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

$$P(\max_{g \in \mathcal{G}_{+,s}} |\hat{\mu}_g^2 - \mu_g^2| > \epsilon) = P(\exists g \in \mathcal{G}_{+,s}, |\hat{\mu}_g^2 - \mu_g^2| > \epsilon)$$

$$= P\left(\bigcup_{g \in \mathcal{G}_{+,s}} \{ |\hat{\mu}_g^2 - \mu_g^2| > \epsilon \} \right)$$

$$\leq \sum_{g \in \mathcal{G}_{+,s}} P( |\hat{\mu}_g^2 - \mu_g^2| > \epsilon)$$

$$\leq G_1 \max_{g \in \mathcal{G}_{+,s}} P( |\hat{\mu}_g^2 - \mu_g^2| > \epsilon)$$

$$= G_1 \max_{g \in \mathcal{G}_{+,s}} P\left( \left| 2\rho_g Z'_g v_g + \frac{Z'_g v_g}{\sqrt{Z'_g Z_g}} \right|^2 > \epsilon \right)$$

$$\leq G \max_{g \in \mathcal{G}_{+,s}} P\left( |2\rho_g Z'_g v_g| > \epsilon \right)$$
Since \( v_{ig} \sim \mathcal{N}(0, \sigma_v^2) \), apply the tail bound for Gaussian random variable,

\[
P \left( |2 \rho_g Z'_g v_g| > \epsilon \right) = P \left( \left| \frac{1}{\sigma_v \sqrt{k_g \sqrt{m_g}}} Z'_g v_g \right| > \frac{\epsilon}{2 \rho_g \sqrt{m_g \sigma_v \sqrt{k_g}}} \right) \leq 2 \exp \left( - \frac{\epsilon^2}{2 \rho_g^2 m_g \sigma_v^2 k_g} \right)
\]

Under Assumption 1, \( k_g < \bar{k} < \infty, \rho_g < \bar{\rho} < \infty \) for \( g \in G_{+s} \) and \( \bar{c} \leq n_g \leq \bar{c} \),

\[
P(\max_{g \in G_{+s}} |\hat{\mu}_g^2 - \mu_g^2| > \epsilon) \lesssim G \exp\left( - \frac{1}{2} \frac{\epsilon^2}{4 \bar{\rho}^2 \bar{c} \bar{N} \sigma_v^2 k} \right) \to 0
\]

The last result holds by picking \( \epsilon = C_1 \sqrt{\frac{N}{G}} \log G \), for \( C_1 > 8 \bar{c} \bar{\rho}^2 \sigma_v^2 \bar{k} \).

Therefore, with probability goes to one,

\[
\min_{g \in G_{+s}} \hat{\mu}_g^2 \geq \min_{g \in G_{+s}} \mu_g^2 - C_1 \sqrt{\frac{N}{G}} \log G \iff \\
\frac{\min_{g \in G_{+s}} \hat{\mu}_g^2}{\sqrt{\frac{N}{G}} \log G} \geq \frac{\min_{g \in G_{+s}} \mu_g^2}{\sqrt{\frac{N}{G}} \log G} - C_1 \to +\infty
\]

The last result is due to \( \min_{g \in G_{+s}} \mu_g^2 \gtrsim \frac{N}{G} \) as a result of Lemma A5.

To prove ii), we have

\[
P(\max_{g \in G_{+s}} |\hat{\mu}_g^2| > \eta) \leq G \max_{g \in G_{+s}} P \left( \left| \frac{Z'_g v_g}{\sigma_v \sqrt{Z'_g Z_g}} \right| > \frac{\eta}{\sigma_v} \right) \leq G \exp\left( - \frac{1}{2} \frac{\eta^2}{\sigma_v^2} \right) \to 0
\]

The last result holds by picking \( \eta = C_2 \log G \) for \( C_2 > 2 \sigma_v^2 \). Therefore,

\[
P(\max_{g \in G_{+s}} |\hat{\mu}_g^2| > C_2 \log G) \to 0 \iff \\
P \left( \frac{\max_{g \in G_{+s}} |\hat{\mu}_g^2|}{\sqrt{\frac{N}{G}} \log G} > C_2 \frac{\sqrt{\log G}}{\sqrt{\frac{N}{G}}} \right) \to 0
\]

which implies \( 0 \leq \frac{\max_{g \in G_{+s}} |\hat{\mu}_g^2|}{\sqrt{\frac{N}{G}} \log G} \leq C_2 \frac{\sqrt{\log G}}{\sqrt{\frac{N}{G}}} \to 0 \), with probability approaching one. Hence,

\[
\frac{\max_{g \in G_{+s}} |\hat{\mu}_g^2|}{\sqrt{\frac{N}{G}} \log G} = o_p(1).
\]
## Appendix B: Additional Simulation Results

### Table A1: Standard Deviation of Existing Estimators

<table>
<thead>
<tr>
<th>$G_{i,s}/G = 0.1$</th>
<th>$G_{i,s}/G = 0.2$</th>
<th>$G_{i,s}/G = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_{pool}$</td>
<td>$\hat{\beta}_{int}$</td>
<td>$\hat{\beta}_{sel.pool}$</td>
</tr>
<tr>
<td>$\rho_{uv}$ &amp; $\rho_{uv} = 0.6, G = 10$ &amp; $\rho_{uv} = 0.8, G = 10$</td>
<td>$\rho_{uv} = 0.6, G = 50$</td>
<td>$\rho_{uv} = 0.8, G = 50$</td>
</tr>
</tbody>
</table>

Note: The table reports the standard deviation of the different estimators among 1000 simulations. The data generating process is $X_{ig}, Z_{ig} \sim \text{i.i.d. } N(0,1)$, $(u_{ig}, v_{ig}) \sim N(0,0,\sigma^2)$, $(1, \rho_{uv}, \rho_{uv})$, $W_{ig} = \rho_{g} Z_{ig} + X_{ig} + v_{ig}$, $Y_{ig} = \beta W_{ig} + X_{ig} + u_{ig}$ for $i = 1, 2, \ldots, n$, where $\beta = 0$, $\rho_g = 0.5$ for $g = 1, \ldots, G_{i,s}$ and $\rho_g = 0$ for $g > G_{i,s}$.

### Table A2: Bias of Existing Estimators

<table>
<thead>
<tr>
<th>$G_{i,s}/G = 0.1$</th>
<th>$G_{i,s}/G = 0.2$</th>
<th>$G_{i,s}/G = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_{pool}$</td>
<td>$\hat{\beta}_{int}$</td>
<td>$\hat{\beta}_{sel.pool}$</td>
</tr>
<tr>
<td>$\rho_{uv}$ &amp; $\rho_{uv} = 0.6, G = 10$ &amp; $\rho_{uv} = 0.8, G = 10$</td>
<td>$\rho_{uv} = 0.6, G = 50$</td>
<td>$\rho_{uv} = 0.8, G = 50$</td>
</tr>
</tbody>
</table>

Note: The table reports the bias of the different estimators among 1000 simulations. The data generating process is $X_{ig}, Z_{ig} \sim \text{i.i.d. } N(0,1)$, $(u_{ig}, v_{ig}) \sim N(0,0,\sigma^2)$, $(1, \rho_{uv}, \rho_{uv})$, $W_{ig} = \rho_{g} Z_{ig} + X_{ig} + v_{ig}$, $Y_{ig} = \beta W_{ig} + X_{ig} + u_{ig}$ for $i = 1, 2, \ldots, n$, where $\beta = 0$, $\rho_g = 0.5$ for $g = 1, \ldots, G_{i,s}$ and $\rho_g = 0$ for $g > G_{i,s}$.
References


