Inference with Many Weak Instruments

Anna Mikusheva[1] and Liyang Sun[2]

Abstract

We develop a concept of weak identification in linear IV models in which the number of instruments can grow at the same rate or slower than the sample size. We propose a jackknifed version of the classical weak identification-robust Anderson-Rubin (AR) test statistic. Large-sample inference based on the jackknifed AR is valid under heteroscedasticity and weak identification. The feasible version of this statistic uses a novel variance estimator. The test has uniformly correct size and good power properties. We also develop a pre-test for weak identification that is related to the size property of a Wald test based on the Jackknife Instrumental Variable Estimator (JIVE). This new pre-test is valid under heteroscedasticity and with many instruments.

`Key words:` instrumental variables, weak identification, dimensionality asymptotics.

`JEL classification codes:` C12, C36, C55.

1 Introduction

Recent empirical applications of instrumental variables (IV) estimation often involve many instruments that together may or may not be strongly relevant. A prominent example is Angrist and Krueger (1991), which started the weak IV literature, uses 180 instruments by interacting dummies for the quarter of birth with state and year of birth. Other examples include papers that employ an empirical strategy known as "judge design" (Maestas et al., 2013; Sampat and Williams, 2015; Dobbie et al., 2018). Fueled by rich administrative data, these papers use the exogenous assignment of cases to judges as instruments for treatment. Since each judge can only process a certain number of cases out of the total court cases, the number of judges (the number of instruments) is usually proportional to the sample size. Another example is the famous Fama-MacBeth procedure in Asset

[1]Department of Economics, M.I.T. Address: 77 Massachusetts Avenue, E52-526, Cambridge, MA, 02139. Email: amikushe@mit.edu. National Science Foundation support under grant number 1757199 is gratefully acknowledged. We are grateful to Josh Angrist, Kirill Evdokimov and Whitney Newey for advice, to Brigham Frandsen for sharing code for simulations, and to Ben Deaner and Sylvia Klosin for research assistance. Replication code is available at `http://economics.mit.edu/grad/lsun20/`.

[2]Department of Economics, M.I.T. Address: 77 Massachusetts Avenue, E52-300, Cambridge, MA, 02139. Email: lsun20@mit.edu.

Pricing (Fama and MacBeth, 1973; Shanken, 1992), which is equivalent to IV estimation procedure with the number of instruments proportional to the number of assets.

This paper answers three questions in an environment with many instruments: how to define weak identification, what to do if identification is weak, and how to pre-test for weak instruments. We model many-instrument asymptotics by allowing the number of instruments to grow at most proportionally with the sample size. Firstly, we define weak identification for linear IV models with many instruments by providing necessary and sufficient conditions for the existence of a consistent test. Secondly, we introduce a test that works when there are many instruments, but is also robust to weak identification and heteroscedasticity. Finally, we propose a pre-test for weak identification. This pre-test forms the basis for a two-step procedure that is analogous to that of Stock and Yogo (2005). The two-step test controls size distortion under many-instrument asymptotics, regardless of the strength of identification or the presence of heteroscedasticity.

For one of our main results, we define weak identification as a situation where an analog of the concentration parameter divided by the *square root of the number of instruments* stays bounded in large samples. We prove that even in a homoscedastic model with known covariance, an asymptotically consistent test does not exist if the ratio of the concentration parameter over the square root of the number of instruments stays bounded in large samples. Thus, a necessary condition for a consistent test to exist is that the concentration parameter grows faster than the square root of the number of instruments. Later, we show that this is also a sufficient condition by constructing a robust test that becomes consistent when this condition is satisfied.

We propose a new jackknifed version of the Anderson-Rubin (AR) test which is robust to both weak identification and heteroscedasticity in a model with many instruments. The new test uses an asymptotic approximation based on a Central Limit Theorem (CLT) for quadratic forms. The new AR test has the correct size regardless of identification strength and becomes consistent as soon as the concentration parameter grows faster than the square root of the number of instruments.

As an important technical contribution, we introduce a novel variance estimator for the quadratic form CLT. The target variance is a quadratic form of the individual (het-

eroscedastic) variances of errors. We apply cross-fitting (Newey and Robins, 2018; Kline et al., 2019) to produce unbiased proxies for the individual variances of errors. We adjust the quadratic form to remove the bias due to correlations between proxies. We prove the consistency of the new estimator under the null and local alternatives.

Finally, we propose a new pre-test for weak identification which is easy to use and is consistent with our definition of weak identification. An empirical researcher can use our pre-test to decide between employing our jackknife AR test if the pre-test suggests that the identification is weak or a Wald test based on the Jackknife Instrumental Variable Estimator (JIVE, Angrist et al., 1999) if the pre-test suggests that the identification is strong. We guarantee the size of this two-step procedure. Chao et al. (2012) prove that JIVE is consistent in a heteroscedastic model when the concentration parameter grows faster than the square root of the number of instruments. Chao et al. (2012) also derive a consistent estimator of the JIVE standard error. The two-step procedure is appealing because when identification is strong, the JIVE-Wald is more efficient and easy to implement and report.

Our pre-test is in the spirit of Stock and Yogo (2005), but it differs from theirs in two important ways. Firstly, our pre-test allows for a general form of heteroscedasticity, while the pre-test proposed in Stock and Yogo (2005) works only under conditionally homoscedastic errors. Secondly, the Stock and Yogo (2005) pre-test is designed for a small number of instruments and is based on the Two-Stage Least Squares (TSLS) estimator. With many instruments TSLS is consistent only when the concentration parameter grows faster than the number of instruments, which makes the Stock and Yogo (2005) pre-test not very informative.

We apply our pre-test to Angrist and Krueger (1991) and find that their identification is strong. Consequently the JIVE confidence set is reliable (has coverage within 5% tolerance level of the declared coverage). Our weak identification-robust jackknife AR confidence set is somewhat wider than the JIVE confidence set but is still informative.

Relation to the Literature. Our paper contributes to both the literature on weak IV and the literature on many instruments. The weak IV literature relates identification strength to the size of the concentration parameter and proposes robust tests that work

only when there are a small number of instruments. Generalizations to many weak instruments either strongly restrict the number of instruments (Andrews and Stock, 2007) or work only under homoscedasticity (Anatolyev and Gospodinov, 2011).

The many instruments literature mostly establishes consistency conditions for particular estimators. For example, Chao and Swanson (2005) show that in a homoscedastic model limited information maximum likelihood (LIML) and bias-corrected TSLS (BTSLS) are consistent when the concentration parameter grows faster than the square root of the number of instruments. In a heteroscedastic model, consistency of LIML and BTSLS requires that the concentration parameter grows faster than the number of instruments. By contrast, JIVE remains consistent when the concentration parameter grows faster than the square root of the number of instruments (Chao et al. (2012)).

The remainder of this paper is organized as follows. In Section 2 we introduce our definition of weak identification in an environment with many instruments. In Section 3 we construct the jackknife AR test and establish its power properties. In Section 4 we present the pre-test and prove that it controls size. Section 5 reports our pre-test results for Angrist and Krueger (1991) and conducts a simulation exercise inspired by Angrist and Frandsen (2019), and Section 6 concludes. Some proofs and additional results may be found in the Supplementary Appendix.

2 Weak Identification with Many Instruments

We study the linear IV regression with a scalar outcome Y_i, a potentially endogenous scalar regressor X_i and a $K \times 1$ vector of instrumental variables Z_i:

$$\begin{cases} Y_i = \beta X_i + e_i, \\ X_i = \Pi_i + v_i, \end{cases} \tag{1}$$

for $i = 1, ..., N$. We denote $\Pi_i = \mathbb{E}[X_i|Z_i]$ and allow the instruments to affect the endogenous regressor in a non-linear way. All results in this paper hold conditionally on a realization of the instruments. Thus, we treat the instruments as fixed (non-random) and Π_i as some constants. The mean-zero errors (e_i, v_i) are independent across i but not

identically distributed and may be heteroscedastic. We assume without loss of generality that there are no controls included in our model as they may be partialled out.

Weak identification under small K is studied extensively in the weak IV literature. For Gaussian homoscedastic errors (e_i, v_i) and linear first stage $(\Pi_i = \pi' Z_i)$, the strength of the instruments corresponds directly to the concentration parameter, $\frac{\pi' Z' Z \pi}{\sigma_v^2}$, where $\sigma_v^2 = Var(v_i)$. The concentration parameter equals the signal-to-noise ratio in the first-stage regression and is related to the bias of the TSLS estimator and the quality of Gaussian approximation for the TSLS t-statistic. For the general case with homoscedastic errors, Staiger and Stock (1997) introduced weak instrument-asymptotics in which one considers a sequence of models so that the concentration parameter converges to a constant as $N \to \infty$. Under this asymptotic embedding, neither a consistent estimator of β nor a consistent test of the null hypothesis that β equals some scalar exists, and the test based on the TSLS t-statistic severely over-rejects.

The magnitude of the concentration parameter is not a good indicator of identification strength when the number of instruments is large. We model large K by considering $K \to \infty$ as $N \to \infty$, with the only restriction that K is at most a fraction of N. Under this many instrument-asymptotics, Theorem 1 below shows that the re-scaled concentration parameter $\frac{\pi' Z' Z \pi}{\sigma_v^2 \sqrt{K}}$ provides a characterization of weak identification in terms of the consistency of tests.

Theorem 1 *Assume we have a sample from model (1) with linear first stage* $\Pi_i = \pi' Z_i$, *where the errors* (e_i, v_i) *are independently drawn from a Gaussian distribution* $\mathcal{N}(0, \Omega)$ *with a known covariance* Ω. *Assume that the* $K \times K$ *matrix* $Z'Z$ *has rank* K *and* $K \to \infty$ *as* $N \to \infty$. *For any sample of size* N *let* Ψ_N *be the class of all tests of size* α *for testing the hypothesis* $H_0 : \beta = \beta_0$, *that is, any* $\psi \in \Psi_N$ *is a measurable function from* $\{(Y_i, X_i, Z_i), i = 1, ..., N\}$ *to the interval* $[0, 1]$ *such that* $\mathbb{E}_{\beta_0, \pi} \psi \leq \alpha$ *for any value of* $\pi \in \mathbb{R}^K$. *Then for any* $\beta^* \neq \beta_0$ *we have*

$$\limsup_{N \to \infty} \max_{\psi \in \Psi_N} \left(\min_{\pi : \frac{\pi' Z' Z \pi}{\sigma_v^2 \sqrt{K}} \leq C} \mathbb{E}_{\beta^*, \pi} \psi \right) < 1.$$

The setting considered in Theorem 1 is quite favorable: the first stage is linear, errors

are Gaussian and homoscedastic with known covariance matrix. So the only unknown parameters are β and π. Theorem 1 states that even in this favorable setting there exists no test that consistently differentiates any β^* from β_0 if the ratio $\frac{\pi' Z' Z \pi}{\sigma_v^2 \sqrt{K}}$ is bounded. Indeed, for any test ψ we can find its guaranteed power $\mathbb{E}_{\beta^*,\pi}\psi$ by minimizing over the alternatives (β^*, π) with bounded $\frac{\pi' Z' Z \pi}{\sigma_v^2 \sqrt{K}}$. We show that even the test that achieves the maximum guaranteed power has guaranteed power strictly less than one asymptotically. Later we show that in a more general heteroscedastic model we can construct a robust test that becomes consistent when $\frac{\Pi' \Pi}{\sqrt{K}} \to \infty$.

Theorem 1 can also be used to characterize weak identification in terms of consistent estimation since it implies there exists no consistent estimator for β when $\frac{\pi' Z' Z \pi}{\sqrt{K}}$ is bounded. Our result complements the literature on estimation with many instruments. Chao and Swanson (2005) show that with homoscedastic errors, when K grows proportionally to the sample size the TSLS estimator is consistent only if $\frac{\pi' Z' Z \pi}{K} \to \infty$, while LIML and BTSLS estimators are consistent when $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$. However, under heteroscedasticity, even when $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$, LIML and BTSLS become inconsistent, but JIVE is still consistent, according to Chao et al. (2012).

The proof of Theorem 1 builds on several classical papers. Following the approach of Andrews et al. (2006), we first reduce the class of tests to those based on a sufficient statistic. Among these tests, the minimal power is achieved by a test invariant to rotations of the instruments. This observation allows us to further reduce our attention to invariant tests, which depend on the data only through its maximal invariant under rotations. Then we derive a limit experiment for $K \to \infty$ similar to that derived in Andrews and Stock (2007). In this limit experiment the minimax power is less than one. Finally we use the argument of Müeller (2011) to bound the desired asymptotic minimax power using the minimax power obtained in the limit experiment.

3 Jackknife AR

The goal of this section is to introduce a test robust to weak identification in the heteroscedastic IV model when the number of instruments, K, is large.

The existing weak IV literature proposes several weak identification-robust tests of the null hypothesis $H_0 : \beta = \beta_0$, when K is small. These tests have correct size when the identification is weak and become consistent when the identification is strong. One example is the Anderson-Rubin (AR) test. Specifically, the IV model (1) implies that under a given null hypothesis $H_0 : \beta = \beta_0$, the exogeneity assumption holds $\mathbb{E}[Z'e(\beta_0)] = 0$ for the implied error $e(\beta_0) = Y - \beta_0 X$. Then under mild assumptions, the scaled sample analog $\frac{1}{\sqrt{N}}Z'e(\beta_0) \Rightarrow N(0, \Sigma)$ satisfies a K-dimensional Central Limit Theorem. The AR statistic is defined as $\frac{1}{N}e(\beta_0)'Z\widehat{\Sigma}^{-1}Z'e(\beta_0)$, where $\widehat{\Sigma}$ is a consistent estimator of $Var\left(\frac{1}{\sqrt{N}}Z'e\right)$. The AR test rejects the null hypothesis when the AR statistic exceeds the $(1-\alpha)$ quantile of the χ^2_K distribution. The AR test has asymptotically correct size regardless of the value of the first stage coefficients Π_i and is asymptotically consistent when an analog of the concentration parameter grows to infinity.

Generalizing the AR statistic to the large-K setting is challenging for multiple reasons. Firstly, the covariance matrix Σ has dimension $K \times K$. Its consistent estimation is problematic if not impossible under general heteroscedasticity. Secondly, the AR statistic under the null has an improperly centered limit distribution because χ^2_K has a very large mean. Thirdly, the K-dimensional Central Limit Theorem provides a poor approximation to the AR statistic when K is large.

We propose an analog of the AR test that is heteroscedasticity-robust and weak identification-robust in the presence of a large number of instruments. Denote the projection matrix $P = Z(Z'Z)^{-1}Z'$. Our test rejects the null of $H_0 : \beta = \beta_0$ when the jackknife AR statistic

$$AR(\beta_0) = \frac{1}{\sqrt{K}\sqrt{\widehat{\Phi}}} \sum_{i=1}^{N} \sum_{j \neq i} P_{ij} e_i(\beta_0) e_j(\beta_0) \tag{2}$$

exceeds the $(1-\alpha)$ quantile of the standard normal distribution. We defer the discussion of the estimator of the variance $\widehat{\Phi}$ to the next subsection.

To address the challenges with the existing AR statistic, the AR statistic we propose uses the default homoscedasticity-inspired weighting $(Z'Z)^{-1}$ in place of $\widehat{\Sigma}^{-1}$. With the $(Z'Z)^{-1}$ weighting, the existing AR statistic has a quadratic form $e(\beta_0)'Pe(\beta_0)$. However, this quadratic form is not centered at zero as it contains the term $\sum_{i=1}^{N} P_{ii}e_i^2$, and each

summand has positive mean. We thus remove this term from the quadratic form. This re-centering can be referred to as leave-one-out or jackknife. In the context of consistent estimation under many instruments, this leave-one-out idea was introduced by Angrist et al. (1999) and fruitfully exploited in a number of papers including Hausman et al. (2012) and Chao et al. (2012). In order to create a test of correct size based on our AR statistic, we use a Central Limit Theorem for quadratic forms proved in Chao et al. (2012) that is restated below.

Assumption 1 *Assume P is an $N \times N$ projection matrix of rank K, $K \to \infty$ as $N \to \infty$ and there exists a constant δ such that $P_{ii} \leq \delta < 1$.*

Lemma 1 *(Chao et al. (2012)) Let Assumption 1 hold for matrix P. Assume the errors η_i are independent, $\mathbb{E}\eta_i = 0$, and there exists a constant C such that $\max_i \mathbb{E}\eta_i^4 < C$, then*

$$\frac{1}{\sqrt{K}\sqrt{\Phi}} \sum_{i=1}^{N} \sum_{j \neq i} P_{ij}\eta_i\eta_j \Rightarrow \mathcal{N}(0,1),$$

where $\Phi = \frac{2}{K}\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}^2 Var(\eta_i)Var(\eta_j)$.

The assumption $P_{ii} \leq \delta < 1$ implies that $\frac{K}{N} = \frac{1}{N}\sum_{i=1}^{N} P_{ii} \leq \delta < 1$. This assumption is often referred to as a balanced design assumption. In the case of group-dummies instruments, P_{ii} is equal to the ratio of the size of the group that observation i belongs to over N. Assumption 1 can be checked for any specific design.

While Lemma 1 requires $K \to \infty$, the Gaussian approximation may work well for smaller K as well. For example, if K is fixed and errors are homoscedastic, then

$$\frac{1}{\sqrt{K}\sqrt{\Phi}} \sum_{i=1}^{N} \sum_{j \neq i} P_{ij}\eta_i\eta_j \Rightarrow \frac{\chi^2_K - K}{\sqrt{2K}} \quad \text{as} \quad N \to \infty.$$

We prove this statement in the Supplementary Appendix S4. While the limit here is not Gaussian it is very well approximated by a standard normal distribution even for relatively small K. The random variable $\frac{\chi^2_K - K}{\sqrt{2K}}$ exceeds the 95% quantile of the standard normal distribution at most 7% of the time for all K, and at most 6% of the time for $K > 40$.

3.1 Variance estimation

In order to conduct asymptotically valid inference based on the normal approximation in Lemma 1, we need an estimator for the scale parameter Φ, which is consistent under the null. One 'naive' estimator that achieves this is $\widehat{\Phi}_1 = \frac{2}{K}\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}^2 e_i^2(\beta_0)e_j^2(\beta_0)$, which uses the square of the implied error as an estimator for the i-th error variance. Under the null when $e_i(\beta_0) = e_i$, the estimator $\widehat{\Phi}_1$ is consistent under relatively mild conditions. However, using $\widehat{\Phi}_1$ in a test would result in poor power. To see this, note that under an alternative value of the parameter $\beta = \beta_0 + \Delta$, we can plug in the first stage and write the implied error $e_i(\beta_0) = Y_i - \beta_0 X_i$ as the sum of a non-trivial mean $\Delta\Pi_i$ and a mean-zero random term $\eta_i = e_i + \Delta v_i$:

$$e_i(\beta_0) = \Delta\Pi_i + \eta_i.$$

While squaring $e_i(\beta_0)$ makes it an unbiased estimator for $Var(e_i)$ under the null, it is biased under the alternative when $\Delta \neq 0$. The bias in $\widehat{\Phi}_1$ grows at the same order as the fourth power of Δ, which brings down the power of the test against distant alternatives.

In order to remove the bias in $e_i^2(\beta_0)$ under the alternatives, one may residualize the implied error before squaring. However, this introduces a bias under the null. Denote $M = I - P$ and let M_i be the ith row of M. Even under the null, the squared residualized error is biased $\mathbb{E}(M_i e)^2 \neq Var(e_i)$. This is because the squared residual contains not only the squared error e_i but also the square of regression estimation mistake. The latter can be large when the number of regressors K is large.

This bias can be removed successfully using the cross-fit variance estimator suggested in Kline et al (2019) and Newey and Robins (2018). Namely, they show that a product of the implied error and residual achieves both goals: it removes the linearly predictable part of the implied error and remains an unbiased estimator of the variance

$$\mathbb{E}\left[\frac{e_i M_i e}{M_{ii}}\right] = Var(e_i).$$

Our challenge is that the scale parameter Φ defined in Lemma 1 is a quadratic form with a double summation. Residuals $M_i e(\beta_0)$ and $M_j e(\beta_0)$ are correlated since they

contain the same estimation mistake. One can show that

$$\mathbb{E}\left[e_i M_i e e_j M_j e\right] = (M_{ii}M_{jj} + M_{ij}^2) Var(e_i) Var(e_j).$$

Our proposed estimator of the scale parameter Φ re-weights each term in the summation to remove the bias described above:

$$\widehat{\Phi} = \frac{2}{K} \sum_{i=1}^{N} \sum_{j \neq i} \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2} \left[e_i(\beta_0) M_i e(\beta_0)\right] \left[e_j(\beta_0) M_j e(\beta_0)\right]. \tag{3}$$

We establish the consistency of $\widehat{\Phi}$ under the null and extend this result to local alternatives.

Assumption 2 *Errors $\epsilon_i, i = 1, ..., N$ are independent with $\mathbb{E}\epsilon_i = 0$, $\max_i \mathbb{E}\|\epsilon_i\|^6 < \infty$, and for some constants c^* and C^* that do not depend on N*

$$c^* \leq \min_i \min_x \frac{x' Var(\epsilon_i) x}{x'x} \leq \max_i \max_x \frac{x' Var(\epsilon_i) x}{x'x} \leq C^*.$$

Theorem 2 *Let Assumption 1 hold for matrix P and Assumption 2 hold for errors e_i, then for $\beta = \beta_0$, we have $\frac{\widehat{\Phi}}{\Phi} \to^p 1$ as $N \to \infty$.*

Theorem 2 combined with Lemma 1 implies that under the null $H_0 : \beta = \beta_0$ our proposed AR statistic has an asymptotically standard normal distribution. Since no assumption about identification is made, the resulting AR test has asymptotically correct size regardless of the strength of identification.

Theorem 3 *Let Assumption 1 hold for matrix P and Assumption 2 hold for errors $\epsilon_i = (e_i, v_i)'$, and $\Pi' M \Pi \leq \frac{C}{K} \Pi' \Pi$. Then for $\beta = \beta_0 + \Delta$ such that $\Delta^2 \cdot \frac{\Pi'\Pi}{K} \to 0$, we have $\frac{\widehat{\Phi}}{\Phi} \to^p 1$ as $N \to \infty$.*

Theorem 3 establishes the consistency of the variance estimator when the null hypothesis does not hold. We use Theorem 3 to derive local power curves of the AR test discussed in the next section. The variance estimator (3) residualizes some implied errors $M_i e(\beta_0)$ to remove non-trivial mean of $e(\beta_0)$ under the alternative. The residualization is complete if the first stage is linear $\Pi_i = \pi' Z_i$. We do not impose such an assumption

in Theorem 3. Instead we require that the approximation of Π_i by a linear combination of instruments improves with the number of instruments as measured by the L_2 norm of the approximation mistake, $\Pi' M \Pi$.

3.2 Power of the Jackknife AR test

Let us introduce a jackknife measure of the information contained in the instruments:

$$\mu^2 = \sum_{i=1}^{N} \sum_{j \neq i} P_{ij} \Pi_i \Pi_j.$$

Theorem 4 *Let $\mathbb{P}_\beta$ be a probability measure describing the distribution of $AR(\beta_0)$ defined in (2) and (3) under model (1) with parameter $\beta = \beta_0 + \Delta$. Assume that the sequence of first stage parameters Π satisfies the following assumptions: $\Pi' M \Pi \leq \frac{C}{K} \Pi' \Pi$ and $\frac{\Pi' \Pi}{K} \to 0$ as $N \to \infty$. If Assumption 1 holds and the errors $\epsilon_i = (e_i, v_i)'$ satisfy Assumption 2, then for any positive constant c we have:*

$$\lim_{N \to \infty} \sup_{|\Delta|^2 \leq c} \sup_{z} \left| \mathbb{P}_\beta \{ AR(\beta_0) < z \} - F\left(z - \frac{\Delta^2 \mu^2}{\sqrt{K\Phi}} \right) \right| = 0, \tag{4}$$

where $F(\cdot)$ is the standard normal cdf. If the sequence of first stage parameters additionally satisfies the condition $\frac{\mu^2}{\sqrt{K\Phi}} \to \infty$, then for any fixed $\Delta \neq 0$ the jackknife AR test is asymptotically consistent:

$$\lim_{N \to \infty} \mathbb{P}_\beta \{ AR(\beta_0) < z \} = 1.$$

Equation (4) of Theorem 4 characterizes the local power curves of the jackknife AR test. The power under the alternative $\beta = \beta_0 + \Delta$ is a function of the distance Δ between the alternative β and the null β_0, the number of instruments K, a measure of identification strength μ^2 and the degree of uncertainty $\sqrt{\Phi}$. Our jackknife AR statistic can be negative, unlike the AR statistic from the small-K case which is always non-negative. We reject the null when $AR(\beta_0)$ exceeds the $(1 - \alpha)$ quantile of the standard normal distribution. Under the alternative $\beta = \beta_0 + \Delta$, the AR statistics has a positive drift which gives rise to

a two-sided test. The second statement of Theorem 4 shows that the AR test consistently distinguishes β from β_0 as long as $\frac{\mu^2}{\sqrt{K}\sqrt{\Phi}} \to \infty$.

Our measure of identification strength, μ^2, has a form similar to the numerator of the concentration parameter defined for the homoscedastic small-K case. Though the two forms are similar, there is an important distinction between them. In our case the signal strength is measured by a jackknife form, while in the homoscedastic small-K case it is measured by $\Pi' P \Pi = \sum_{i=1}^{N} \sum_{j=1}^{N} P_{ij} \Pi_i \Pi_j$. The instruments may affect the endogenous regressor in an arbitrarily non-linear way, and only the projection of Π onto the linear space of the instruments is used by the linear IV regression. Thus the projection matrix appears naturally in our measure of identification strength. If the effect of instruments on the regressor (Π) is well approximated by the linear first stage ($\Pi' M \Pi \leq \frac{C}{K} \Pi' \Pi$), then the strength of identification has the same order as $\Pi'\Pi$ in the sense that they grow to infinity or stay bounded simultaneously. Indeed, under Assumption 1 we have:

$$\left(1 - \delta - \frac{C}{K}\right) \Pi'\Pi \leq \mu^2 = \Pi'\Pi - \Pi' M \Pi - \sum_{i=1}^{N} P_{ii} \Pi_i^2 \leq \Pi'\Pi.$$

Theorem 4 implies that $\frac{\mu^2}{\sqrt{K}} \to \infty$ is a sufficient condition for the consistency of the jackknife AR test. When the first stage is well approximated by linear combination of the instruments, this translates to a sufficient condition of $\frac{\Pi'\Pi}{\sqrt{K}} \to \infty$. This compliments Theorem 1 which implies that $\frac{\Pi'\Pi}{\sqrt{K}} \to \infty$ is necessary for the consistency of any test. It is worth noticing that the condition $\frac{\Pi'\Pi}{K} \to 0$ imposed by Theorem 4 is quite weak as it covers both weakly and strongly identified cases.

4 Pre-test for Weak Identification

In a prominent paper, Stock and Yogo (2005) introduced a pre-test for weak identification that has gained enormous popularity in applied work. In homoscedastic IV models with small K, the concentration parameter fully characterizes the worst bias of the TSLS as a fraction of the OLS bias and the worst rejection rate of TSLS-Wald test. Since the first stage F statistic measures the concentration parameter, Stock and Yogo (2005) suggest

a set of cut-offs for the first stage F statistic, above which a researcher can guarantee with high (prespecified) probability that the bias of TSLS is not larger than 10% of the OLS bias, or that the TSLS-Wald statistic does not over-reject by more than 5%. The cut-offs depend on the goal (bias or size) and the number of instruments. However, these details seem to be mostly disregarded in empirical practice, as the most common guidance suggests a cut-off of 10, regardless of the goal or the number of instruments.

As with any procedure of such generality, it suffers from multiple drawbacks. First, the pre-test is valid only if the model is homoscedastic. Andrews (2018) shows that in models calibrated to commonly-used data sets with heteroscedasticity one may find cases with the first stage F statistics exceeding 1000, that have large over-rejections of the TSLS-Wald test.

Second, the TSLS estimator is less robust to weak identification when K is large. In a homoscedastic model when K is growing proportionally to the sample size, the TSLS estimator is consistent only if $\frac{\pi' Z' Z \pi}{K} \to \infty$, while LIML and BTSLS estimators are consistent when $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$ (see Chao and Swanson (2005)). In this case, the pre-test becomes too conservative. Indeed, if $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$ but $\frac{\pi' Z' Z \pi}{K} \nrightarrow \infty$, then the pre-test most likely declares weak identification as the expectation of the first stage F equals to $\frac{\pi' Z' Z \pi}{K \sigma_v^2} + 1$, even though there exist consistent estimators and a reasonable Wald-test can be constructed.

We propose a new pre-test for weak identification that allows us to form a two-step procedure: a researcher first assesses instrument strength based on our pre-test and then uses the JIVE-Wald test if the instruments appear strong and our jackknife AR test if they appear weak. We can guarantee the size of such two-step procedure in a heteroscedastic IV model with large K. Our pre-test uses an empirical measure of $\frac{\mu^2}{\sqrt{K}}$, whose value characterizes weak identification as discussed in the previous sections:

$$\widetilde{F} = \frac{1}{\sqrt{K}\sqrt{\widehat{\Upsilon}}} \sum_{i=1}^{N} \sum_{j \neq i} P_{ij} X_i X_j, \tag{5}$$

here $\widehat{\Upsilon} = \frac{2}{K} \sum_i \sum_{j \neq i} \frac{P_{ij}^2}{M_{ii} M_{jj} + M_{ij}^2} X_i M_i X X_j M_j X$ is an estimate of the variance Υ defined

in (12). The JIVE-Wald test uses the JIV2 estimator introduced in Angrist et al. (1999):

$$\widehat{\beta}_{JIVE} = \frac{\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}Y_iX_j}{\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}X_iX_j}.$$

Our choice of JIVE is based on two considerations. First, according to Hausman et al. (2012), in a heteroscedastic IV model, when $\frac{\pi' Z' Z \pi}{\sqrt{K}} \to \infty$, LIML and BTSLS become inconsistent, but JIVE is consistent. Second, the JIVE estimator is a ratio of two quadratic forms similar to the jackknife AR statistic. We use the following estimator of the JIVE variance, that is a cross-fit version of the estimator derived in Chao et al. (2012):

$$\widehat{V} = \frac{\sum_{i=1}^{N}\left(\sum_{j\neq i} P_{ij}X_j\right)^2 \frac{\widehat{e}_i M_i \widehat{e}}{M_{ii}} + \sum_{i=1}^{N}\sum_{j\neq i} \widetilde{P}_{ij}^2 M_i X \widehat{e}_i M_j X \widehat{e}_j}{\left(\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}X_iX_j\right)^2},$$

where $\widehat{e}_i = Y_i - X_i\widehat{\beta}_{JIVE}$. The Wald statistic is defined as $Wald(\beta_0) = \frac{\left(\widehat{\beta}_{JIVE}-\beta_0\right)^2}{\widehat{V}}$.

Theorem 5 *Let Assumption 1 hold for matrix P and Assumption 2 hold for errors $\epsilon_i = (e_i, v_i)'$. Assume that $\Pi' M \Pi \leq \frac{C\Pi'\Pi}{K}$ and $\frac{\Pi'\Pi}{K^{2/3}} \to 0$ as $N \to \infty$. Then for $\beta = \beta_0$,*

$$\left(Wald(\beta_0), \widetilde{F}\right) \Rightarrow \left(\frac{\xi^2}{1 - 2\varrho\frac{\xi}{\nu} + \frac{\xi^2}{\nu^2}}, \nu\right), \tag{6}$$

where ξ and ν are two normal random variables with means 0 and $\frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}}$, unit variances and correlation coefficient ϱ defined in equation (12).

Theorem 5 shows that the distribution of the JIVE-Wald statistics can be quite different from its conventional χ_1^2 limit when $\frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}}$ is small. If $\frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}}$ is large, then most realizations of the random variable ν are large as well and the limit of the JIVE-Wald is close to the distribution of ξ^2, which is χ_1^2. This suggests that $\frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}}$ is a good measure for identification strength. We notice that the limit expression for the JIVE-Wald statistics is similar to the limit distribution derived by Stock and Yogo (2005, formula (2.22)) for TSLS-Wald in homoscedastic weak IV with small K.

Using Theorem 5 we can calculate the worst asymptotic rejection rate of the JIVE-

Wald test as a function of $\frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}} = x$:

$$R_{\alpha}^{\max}(x) = \max_{\varrho \in [-1,1]} \mathbb{P}_{x,\varrho} \left\{ \frac{\xi^2}{1 - 2\varrho\frac{\xi}{\nu} + \frac{\xi^2}{\nu^2}} \geq \chi^2_{1,1-\alpha} \right\},$$

where $\mathbb{P}_{x,\varrho}$ is the probability distribution of ξ, ν described in Theorem 5. For a typical test with nominal size $\alpha = 5\%$, we find that $\frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}} = x > 2.5$ implies $R_{5\%}^{\max}(x) < 10\%$. Theorem 5 also allows us to construct a 5%-test for the null hypothesis that the unknown strength of identification parameter $\frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}}$ is higher than 2.5. This test is based on the statistic $\widetilde{F}$ and accepts whenever $\widetilde{F} > 4.14$. Using Bonferroni bounds we obtain the following statement:

Corollary 1 *Let all assumptions of Theorem 5 hold. Then a two-step test for the null hypothesis $H_0 : \beta = \beta_0$ that accepts the null if $\widetilde{F} > 4.14$ and $Wald(\beta_0) < \chi^2_{1,0.95}$ or if $\widetilde{F} \leq 4.14$ and $AR(\beta_0) < z_{0.95}$, has an asymptotic size smaller than 15%.*

The pre-test we propose is to compare $\widetilde{F}$ with the cut-off of 4.14. If $\widetilde{F}$ exceeds the cut-off one may proceed using JIVE test/confidence set, otherwise one is advised to employ weak-identification robust jackknife AR test. The attraction of the two-step procedure is that confidence sets based on the JIVE-Wald test is relatively easy to construct and is well understood by the practitioners. As we illustrate in simulations, the Jackknife AR confidence sets tend to be wider than the JIVE-Wald confidence sets when identification is strong. Simulations also suggest the Bonferroni bounds derived in Corollary 1 tend to be conservative, as the actual size of the two-step test does not exceed 7%.

5 Empirical Illustration: Return to Education

Angrist and Krueger (1991) (AK91 in what follows) provided a motivating example for the weak identification literature, starting with the seminal work by Bound et al. (1995). Staiger and Stock (1997) suggested that the relatively low value of the first stage F statistic can be seen as a sign of potential weak instruments in the AK91 application. Hansen et al. (2008) argued that "many instruments" may be a more relevant description

	FF	$\widetilde{F}$	JIVE-Wald	Jackknife AR
180 instruments	2.428	13.422	[0.066,0.132]	[0.008,0.201]
1530 instruments	1.27	6.173	[0.024,0.121]	[-0.047, 0.202]

Table 1: AK91 Pre-test Results

Notes: Results on pre-tests for weak identification and confidence sets for IV specification underlying Table VII Column (6) of Angrist and Krueger (1991) using the original data. FF is the first stage F statistic of Stock and Yogo (2005), $\widetilde{F}$ is the statistic introduced in (5). The JIVE-Wald confidence set is described in Section 4. The jackknife AR confidence set is based on analytical test inversion.

of the identification issue encountered in AK91, as instruments are possibly not weak collectively. They suggested that estimators other than TSLS may restore the accuracy for standard inferences. We apply our proposed pre-test statistics $\widetilde{F}$ to the original AK91 application to assess whether instruments are weak given that there are many of them.

The original AK91 application estimated the effect of schooling (X_i) on log weekly wage (Y_i) using quarter of birth as instruments in a sample of 329,509 men born 1930-39 from the 1980 census.[3] There are multiple specifications in the original AK91 study. We focus on the specification with 180 instruments and also an extension of this specification using 1530 instruments. The 180 instruments include 30 quarter and year of birth interactions (QOB-YOB) and 150 quarter and state of birth interactions (QOB-POB). For the second specification with 1530 instruments, we also include full interactions among QOB-YOB-POB. Table 1 reports the first stage F statistics (FF), our proposed pre-test statistics $\widetilde{F}$ introduced in (5), confidence sets based on the JIVE-Wald and jackknife AR statistics. While the first stage F statistic is below 10 and the pre-test from Stock and Yogo (2005) would point toward weak identification for both specifications, the instruments turn out to be strong in both specifications based on our pre-test. As a result, the reported confidence sets based on a norminal 5% JIVE-Wald test are reliable, as the actual size is at most 15%. The confidence sets based on our jackknife AR statistic are wider, yet still informative.

[3]With this sample size, we cannot vectorize calculations involving P_{ij}^2 (for jackknife AR and pre-test) due to memory constraint. However, it is still relatively fast to execute the non-vectorized code, which takes around 20 minutes.

N	K	Avg. $\widetilde{F}$	OLS bias	2SLS bias	2SLS size	LIML bias	LIML size	JIVE bias	JIVE size
4,923	154	4.99	0.26	0.17	96.6%	-0.001	0.6%	-0.03	5%
3,209	135	3.35	0.26	0.19	95.7%	-0.05	2.7%	-0.06	5.2%
1,599	111	1.77	0.26	0.21	92.3%	-0.89	14.5%	1.22	3.6%

Table 2: AK91 Simulation Results: Bias of different estimators and Size of Non-robust Tests

5.1 Monte Carlo Experiments

Through Monte Carlo simulations we show that the jackknife AR and the pre-test we develop are robust to many weak instruments unlike canonical IV estimators. To illustrate the practical importance of many weak instruments, we attempt to preserve the structure of AK91. Specifically, we adopt the simulation design by Angrist and Frandsen (2019). There is very little endogeneity in the original AK91, which makes it hard to study the biases of different estimators. Thus, we follow Angrist and Frandsen (2019) to introduce additional omitted variable bias to the simulated data. The simulated data has a nonlinear first stage and is heteroscedastic. We deviate from Angrist and Frandsen (2019) in two respects. First, we vary the sample size N of the simulated data to be 1.5%, 1% and 0.5% of the original sample size. This is to vary the identification strength. We report the identification strength by the average $\widetilde{F}$ across simulations. Simulations with sample size equal to 1.5% of the original sample size produce strong identification in our definition, 1% still produce strong identification but close to the weak identification region, while 0.5% produce weak identification.When we reduce the sample size we also need to exclude the instruments of the groups that are no longer populated. Second, both in data simulation and in estimation we do not include controls in order to isolate the implications of many instruments. The Appendix provides more details on our simulation design.

We evaluate the performance of common estimators and tests based on 1000 simulation draws. In Table 2, we report the bias and Wald tests size of OLS, 2SLS, LIML and JIVE estimators. For the Wald test based on the LIML estimator, we calculate the standard errors as in Hansen et al. (2008). They corrected the canonical standard error estimator to be robust to many instruments, but this test is not robust to heteroscedasticity, as LIML itself is inconsistent under heteroscedasticity. For the Wald test based on the

N	K	Avg. FF	Avg. $\widetilde{F}$	jackknife AR	pre-test	two-step test
4,923	154	1.63	4.99	5.1%	70.5%	5.8%
3,209	135	1.44	3.35	5.6%	26.7%	6.6%
1,599	111	1.24	1.77	6.3%	4.5 %	7.2%

Table 3: AK91 Simulation Results: Size of Robust Tests

JIVE estimator, we calculate the heteroscedasticity-robust standard errors as described in Section 4.

We find that due to many instruments 2SLS has large bias even under strong identification. While Hausman et al. (2012) show LIML is inconsistent under many instruments and heteroscedasticity, LIML is not too biased in our simulated data, as long as identification is not weak. We find that JIVE has low bias when identification is strong, but its bias increases when identification is weak. The Wald test based on either LIML or JIVE is not robust to many weak instruments, and we find substantial size distortion for LIML under weak identification. Surprisingly we do not find large size distortion for JIVE.

In Table 3 we report the rejection frequency of the robust test we developed in this paper based on the jackknife AR test statistic. We find that the jackknife AR controls size even under weak identification. Our proposed pre-test also controls size and is able to switch to the JIVE-Wald test when identification is strong. In contrast, the first stage F statistics of Stock and Yogo (2005) (FF) are very small even under strong identification, which makes it not very informative.

Finally, in Table 4 we compare the length of confidence intervals formed by inverting various tests. In particular, when identification is strong, jackknife AR confidence sets are longer (less efficient) but are not unreasonably long compared to the Wald tests based on LIML and JIVE. In this case, a pre-test can improve the efficiency by switching to the Wald test based on JIVE. As with the canonical AR test, the jackknife AR test can result in confidence intervals with infinite length. We report the probability of infinite length in the last column of Table 4, and note that such probability increases as identification gets weaker.

N	K	Avg. $\widetilde{F}$	2SLS	LIML	JIVE	jackknife AR	infinite jackknife AR
4,923	143	4.99	0.18	1.14	0.81	1.66	0.4%
3,209	135	3.35	0.20	1.23	1.41	2.76	14.2%
1,599	111	1.77	0.24	1.46	5244	6.89	51.1%

Table 4: AK91 Simulation Results, Length of Confidence Interval

6 Conclusion

In this paper, we argue that we can characterize weak identification as an environment with many instruments when an analog of the concentration parameter staying bounded relative to the square root of the number of instruments in large samples. We introduce a jackknifed version of the AR test that is robust to our definition of weak identification and heteroscedasticity. We also propose a pre-test for weak identification and correspondingly a two-step testing procedure in the spirit of Stock and Yogo (2005). Unlike the pre-test proposed by Stock and Yogo (2005), our two-step test controls size distortion even under heteroscedasticity and with many instruments. As an empirical example, our pre-test rejects weak identification in Angrist and Krueger (1992) where up to 1530 instruments are used.

7 References

Anatolyev, S., and Gospodinov, N. (2011). "Specification Testing in Models with Many Instruments." *Econometric Theory* 27, 427–441.

Andrews, I. (2018). "Valid Two-Step Identification-Robust Confidence Sets for GMM." *The Review of Economics and Statistics*, 100, 337–348. *Supplementary Appendix*

Andrews, D.W.K., and Stock, J.H. (2007). "Testing with many weak instruments." *Journal of Econometrics* 138, 24–46.

Andrews D, Moreira M, Stock J. (2006). "Optimal two-sided invariant similar tests of instrumental variables regression." *Econometrica* 74:715–752

Angrist, J.D., and Frandsen B. (2019). "Machine Labor" *NBER working paper* 26584.

Angrist, J.D., Imbens, G.W., and Krueger, A.B. (1999). "Jackknife instrumental variables estimation." *Journal of Applied Econometrics* 14, 57–67.

Angrist, J.D., and Krueger, A.B. (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, 979–1014.

Bound, J., Jaeger, D.A. and Baker, R.M. (1995) "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90:430, 443–450.

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

Chao, J.C., Swanson, N.R., Hausman, J.A., Newey, W.K., and Woutersen, T. (2012). "Asymptotic Distribution of JIV in a heteroscedastic IV Regression with Many Instruments." *Econometric Theory* 28, 42–86.

Dobbie, W., Goldin, J., and Yang, C.S. (2018). "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review* 108, 201–240.

Fama, E.F., and MacBeth, J.D. (1973). "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81, 607–636.

Hansen, C., Hausman, J., and Newey, W. (2008). "Estimation With Many Instrumental Variables." *Journal of Business & Economic Statistics* 26, 398–422.

Hausman, J.A., Newey, W.K., Woutersen, T., Chao, J.C., and Swanson, N.R. (2012). "Instrumental variable estimation with heteroscedasticity and many instruments." *Quantitative Economics* 3, 211–255.

Kleibergen F. (2002). "Pivotal statistics for testing structural parameters in instrumental variables regression." *Econometrica* 70:1781–1803.

Kline, P., Saggio, R., and Sølvsten, M. (2018). "Leave-out estimation of variance components."

Maestas, N., Mullen, K.J., and Strand, A. (2013). "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *American Economic Review* 103, 1797–1829.

Müller, U. K. (2011). "Efficient Tests Under a Weak Convergence Assumption." *Econometrica* 79 (2): 395–435.

Newey, W. (2004). "Many Instrument Asymptotics."

Newey, W.K., and Robins, J.R. (2018). "Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation."

Newey, W.K., and Windmeijer, F. (2009). "Generalized Method of Moments With Many Weak Moment Conditions." *Econometrica* 77, 687–719.

Sampat, B., and Williams, H.L. (2019). "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome." *American Economic Review* 109, 203–236.

Shanken, J. (1992). "On the Estimation of Beta-Pricing Models." *The Review of Financial Studies* 5, 1–33.

Staiger, D., and Stock, J.H. (1997). "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–86.

Stock, J.H., and Yogo, M. (2005). "Testing for weak instruments in Linear Iv regression. In Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg," pp. 80–108.

8 Appendix with Proofs

Let C be a universal constant (that may be different in different lines but does not depend on N or K). Denote $\sigma_i^2 = Var(e_i), \varsigma_i^2 = Var(v_i)$, $\gamma_i = cov(e_i, v_i)$, and $\widetilde{P}_{ij}^2 = \frac{P_{ij}^2}{M_{ii}M_{jj}+M_{ij}^2}$.

Proof of Theorem 1. Denote A to be an upper-triangular matrix, such that $A\Omega A' = I_2$. The sufficient statistic in model (1) is

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = (A \otimes I_K) \cdot \begin{pmatrix} (Z'Z)^{-1/2}Z'Y \\ (Z'Z)^{-1/2}Z'X \end{pmatrix} \sim N\left(\begin{pmatrix} \widetilde{\beta}\Pi \\ \Pi \end{pmatrix}, I_{2K} \right) \tag{7}$$

where $\widetilde{\beta} = (1,0)A(\beta,1)'$ is a (known) linear one-to-one transformation of β. Denote the corresponding null and alternative as $\widetilde{\beta}_0$ and $\widetilde{\beta}^*$. We denote also $\Pi = \frac{(Z'Z)^{1/2}\pi}{\sigma_v}$, which is one-to-one transformation of π. It is enough to restrict attention to the tests that depend on the data through sufficient statistics only. Indeed, for any test $\psi \in \Psi_N$ we may construct a test $\psi^S = \mathbb{E}(\psi|\xi_1, \xi_2)$ which depends on the data only through the sufficient

statistics. Due to the law of iterated expectations the size and the power of ψ^S is the same as the initial ψ.

Let $\mathcal{U}$ be the group of rotations on $\mathbb{R}^K$, that is $U \in \mathcal{U}$ are such $U'U = I_K$. Notice that the model is invariant to group $\mathcal{U}$, namely if (ξ_1, ξ_2) satisfy model (7) with parameters $(\widetilde{\beta}, \Pi)$ then $(U\xi_1, U\xi_2)$ satisfy model (7) with parameters $(\widetilde{\beta}, U\Pi)$. Note that $\Pi'\Pi = (U\Pi)'(U\Pi)$. This implies that for any function f we have $\mathbb{E}_{(\widetilde{\beta},\Pi)} f(U\xi_1, U\xi_2) = \mathbb{E}_{(\widetilde{\beta},U\Pi)} f(\xi_1, \xi_2)$.

We call a test $\psi = \psi(\xi_1, \xi_2)$ invariant to rotations iff for any $U \in \mathcal{U}$ we have $\psi(U\xi_1, U\xi_2) = \psi(\xi_1, \xi_2)$ for all realizations of (ξ_1, ξ_2). The maximum in Theorem 1 is achieved at an invariant test. Indeed, take any test $\psi \in \Psi_N$ that has size α, that is, $\mathbb{E}_{(\widetilde{\beta}_0,\Pi)} \psi(\xi_1, \xi_2) \leq \alpha$ for all Π. Let us consider a new test $\psi^*(\xi_1, \xi_2) = \int_{U \in \mathcal{U}} \psi(U\xi_1, U\xi_2) dU$, where the integral is taken uniformly over the unit sphere in $\mathbb{R}^K$. By construction, ψ^* is an invariant test as for any $\widetilde{U} \in \mathcal{U}$, we have $U\widetilde{U} \in \mathcal{U}$ for all $U \in \mathcal{U}$ so that

$$\psi^*(\widetilde{U}\xi_1, \widetilde{U}\xi_2) = \int_{U \in \mathcal{U}} \psi(U\widetilde{U}\xi_1, U\widetilde{U}\xi_2) dU = \int_{U \in \mathcal{U}} \psi(U\xi_1, U\xi_2) dU.$$

$$\mathbb{E}_{(\widetilde{\beta}_0,\Pi)} \psi^*(\xi_1, \xi_2) = \int_{U \in \mathcal{U}} \left\{ \mathbb{E}_{(\widetilde{\beta}_0,\Pi)} \psi(U\xi_1, U\xi_2) \right\} dU = \int_{U \in \mathcal{U}} \left\{ \mathbb{E}_{(\widetilde{\beta}_0,U\Pi)} \psi(\xi_1, \xi_2) \right\} dU \leq \alpha.$$

So, it has correct size. Now we check that the minimal power of ψ^* achieved over alternatives $(\widetilde{\beta}^*, \Pi)$ with Π such that $\frac{\Pi'\Pi}{\sqrt{K}} = C$ is not smaller than that of ψ. Assume that the minimum of power for test ψ is achieved at the alternative Π^*: $\min_{\frac{\Pi'\Pi}{\sqrt{K}}=C} \mathbb{E}_{(\widetilde{\beta}^*,\Pi)} \psi(\xi_1, \xi_2) = \mathbb{E}_{(\widetilde{\beta}^*,\Pi^*)} \psi(\xi_1, \xi_2)$. Then, similarly to above:

$$\min_{\frac{\Pi'\Pi}{\sqrt{K}}=C} \mathbb{E}_{(\widetilde{\beta}^*,\Pi)} \psi^*(\xi_1, \xi_2) = \min_{\frac{\Pi'\Pi}{\sqrt{K}}=C} \int_{U \in \mathcal{U}} \left\{ \mathbb{E}_{(\widetilde{\beta}^*,U\Pi)} \psi(\xi_1, \xi_2) \right\} dU \geq$$

$$\geq \int_{U \in \mathcal{U}} \min_{\frac{\Pi'\Pi}{\sqrt{K}}=C} \left\{ \mathbb{E}_{(\widetilde{\beta}^*,U\Pi)} \psi(\xi_1, \xi_2) \right\} dU = \mathbb{E}_{(\widetilde{\beta}^*,\Pi^*)} \psi(\xi_1, \xi_2).$$

All invariant tests depend on the data only through maximal invariant. Thus, we should only consider tests that depend on the data through statistics $Q = (Q_1, Q_2, Q_3) =$

$(\xi_1'\xi_1, \xi_1'\xi_2, \xi_2'\xi_2)$. If $\Pi'\Pi/\sqrt{K} \to C$ then Q converges to the following distribution:

$$\begin{pmatrix} \frac{\xi_1'\xi_1 - K}{\sqrt{2K}} \\ \frac{\xi_1'\xi_2}{\sqrt{K}} \\ \frac{\xi_2'\xi_2 - K}{\sqrt{2K}} \end{pmatrix} \Rightarrow N\left(\begin{pmatrix} \widetilde{\beta}^2 \frac{C}{\sqrt{2}} \\ \widetilde{\beta} C \\ \frac{C}{\sqrt{2}} \end{pmatrix}, I_3 \right) = \begin{pmatrix} Q_{\infty,1} \\ Q_{\infty,2} \\ Q_{\infty,3} \end{pmatrix} = Q_\infty. \tag{8}$$

According to Theorem 1 of Müeller (2011) the limit of the maximal power of tests in experiment based on Q is bounded above by the maximal power achieved in the limit experiment described on Q_∞ as defined in the right hand side of equation (8). Notice that the maximal achievable power $\mathbb{E}_{\widetilde{\beta}^*, C}\psi^*(Q_\infty)$ is strictly less than 1 for any fixed β^* and fixed C. Indeed, the best achievable power in the limit experiment (8) is no more than the best achievable power in the experiment when C is known. If C is known, the optimal test follows from the Neyman-Pearson lemma, and is less than 1.

Proof of Theorem 2. Assumptions 1 and 2 imply

$$1 \geq \frac{1}{K}\sum_i \sum_{j \neq i} P_{ij}^2 = \frac{1}{K}\sum_i \sum_j P_{ij}^2 - \frac{1}{K}\sum_i P_{ii}^2 \geq 1 - \delta \frac{1}{K}\sum_i P_{ii} = 1 - \delta.$$

Thus, $(1-\delta)(c^*)^2 < \Phi < (C^*)^2$ and it is sufficient to prove that $\widehat{\Phi} - \Phi \to^p 0$. The last statement holds due to Lemma 2 applied to $\xi_i = (e_i, e_i, e_i)'$. □

Lemma 2 *Let Assumption 1 hold. Assume the errors* $\xi_i = (\xi_i^{(1)}, \xi_i^{(2)}, \xi_i^{(3)})'$ *are independent mean zero random vectors with* $\max_i \mathbb{E}\|\xi_i\|^6 < C$. *Then as* $N \to \infty$, *we have:*

$$\frac{1}{K}\sum_i \sum_{j \neq i} \left\{ \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2} \left[\xi_i^{(1)} M_i \xi^{(2)}\right] \left[\xi_j^{(1)} M_j \xi^{(3)}\right] - P_{ij}^2 \mathbb{E}\left[\xi_i^{(1)}\xi_i^{(2)}\right] \mathbb{E}\left[\xi_j^{(1)}\xi_j^{(3)}\right] \right\} \to^p 0.$$

Proof of Lemma 2. Notice that

$$\frac{1}{K}\sum_i \sum_{j \neq i} P_{ij}^2 \mathbb{E}\left[\xi_i^{(1)}\xi_i^{(2)}\right] \mathbb{E}\left[\xi_j^{(1)}\xi_j^{(3)}\right] = \frac{1}{K}\sum_i \sum_{j \neq i} \widetilde{P}_{ij}^2 \mathbb{E}\left[\xi_i^{(1)}\xi_i^{(2)}\xi_j^{(1)}\xi_j^{(3)}\right].$$

Define $\xi_{ij} = \xi_i^{(1)} M_i \xi^{(2)} \xi_j^{(1)} M_j \xi^{(3)} - \mathbb{E}\left[\xi_i^{(1)} M_i \xi^{(2)} \xi_j^{(1)} M_j \xi^{(3)}\right]$, then we need to prove that $\frac{1}{K}\sum_i \sum_{j \neq i} \widetilde{P}_{ij}^2 \xi_{ij} \to^p 0$. Since $\frac{1}{K}\sum_i \sum_{j \neq i} \widetilde{P}_{ij}^2 \xi_{ij}$ has zero mean, it is sufficient to show that

the variance of each term in expression (9) defined below converges to zero (here I_4 is a summation over distinct indexes (i, i', j, j')):

$$\mathbb{E}\left(\frac{1}{K}\sum_i\sum_{j\neq i}\widetilde{P}_{ij}^2\xi_{ij}\right)^2 = \frac{1}{K^2}\sum_i\sum_{j\neq i}\widetilde{P}_{ij}^4\mathbb{E}\xi_{ij}^2 + \\ +\frac{1}{K^2}\sum_i\sum_{j\neq i}\sum_{i'\neq\{i,j\}}\widetilde{P}_{ij}^2\widetilde{P}_{ii'}^2\mathbb{E}\xi_{ij}\xi_{ii'} + \frac{1}{K^2}\sum_{I_4}\widetilde{P}_{ij}^2\widetilde{P}_{i'j'}^2\mathbb{E}\xi_{ij}\xi_{i'j'}. \tag{9}$$

First, we prove that $\max_{i,j}\mathbb{E}\xi_{ij}^2 < C$. We expand $\xi_{ij} = A_{1,ij} + A_{2,ij} + A_{3,ij}$, where:

$$A_{1,ij} = M_{ii}M_{jj}\left(\xi_i^{(1)}\xi_i^{(2)}\xi_j^{(1)}\xi_j^{(3)} - \mathbb{E}[\xi_i^{(1)}\xi_i^{(2)}\xi_j^{(1)}\xi_j^{(3)}]\right) + M_{ij}^2\left(\xi_i^{(1)}\xi_i^{(3)}\xi_j^{(1)}\xi_j^{(2)} - \mathbb{E}[\xi_i^{(1)}\xi_i^{(3)}\xi_j^{(1)}\xi_j^{(2)}]\right),$$
$$A_{2,ij} = \xi_i^{(1)}\xi_j^{(1)}\sum_{i'\neq\{i,j\}}\left(M_{ii}M_{ji'}\xi_i^{(2)}\xi_{i'}^{(3)} + M_{ii'}M_{ij}\xi_{i'}^{(2)}\xi_i^{(3)} + M_{jj}M_{ii'}\xi_{i'}^{(2)}\xi_j^{(3)} + M_{ji'}M_{ij}\xi_j^{(2)}\xi_{i'}^{(3)}\right),$$
$$A_{3,ij} = \xi_i^{(1)}\xi_j^{(1)}\sum_{i'\neq\{i,j\}}\sum_{j'\neq\{i,j\}}M_{ii'}M_{jj'}\xi_{i'}^{(2)}\xi_{j'}^{(3)}.$$

It is sufficient to show that $\max_{i,j}\mathbb{E}A_{s,ij}^2$ is bounded for all $s = 1, 2, 3$. The moment condition implies $\mathbb{E}A_{1,ij}^2 \leq C\left(M_{ii}M_{jj} + M_{ij}^2\right)^2 \leq C$. Below we use that non-zero correlations between summands in $A_{s,ij}$ imply that some indexes must coincide. We also use Lemma S1.1 from the Supplementary Appendix:

$$\mathbb{E}A_{2,ij}^2 \leq C\sum_{i'}(M_{ii}M_{ji'} + M_{ii'}M_{ij} + M_{jj}M_{ii'} + M_{ji'}M_{ij})^2 \leq C,$$
$$\mathbb{E}A_{3,ij}^2 \leq C\sum_{i'\neq\{i,j\}}\sum_{j'\neq\{i,j\}}\left(P_{ii'}^2P_{jj'}^2 + |P_{ii'}P_{jj'}P_{ij'}P_{ji'}|\right) \leq C.$$

Next notice that

$$\widetilde{P}_{ij}^2 = \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2} \leq \frac{P_{ij}^2}{(1-P_{ii})(1-P_{jj})} \leq \frac{1}{(1-\delta)^2}P_{ij}^2. \tag{10}$$

Lemma B1 in Chao et al (2012) gives that $\sum_i\sum_{j\neq i}P_{ij}^4 \leq K$ and $\sum_i\sum_{j\neq i}\sum_{j'\neq i,j'\neq j}P_{ij}^2P_{ij'}^2 \leq K$. Thus, given the bound on $\max_{i,j}E\xi_{ij}^2 < C$ and by Cauchy-Schwarz inequality $\max_{i,j,k}|\mathbb{E}\xi_{ij}\xi_{ik}| < C$, the first two terms in expression (9) converge to zero.

For the last term in (9), since i, i', j, j' are all distinct, we have $\mathbb{E}A_{1,ij}A_{s,i'j'} = 0$ for

$s = 2, 3$, and $\mathbb{E}A_{2,ij}A_{3,i'j'} = 0$. The non-zero terms in $\mathbb{E}\xi_{ij}\xi_{i'j'}$ are

$$\begin{aligned}|\mathbb{E}A_{2,ij}A_{2,i'j'}| \leq & C\left|(M_{ii}M_{jj'} + M_{ij}M_{ij'})(M_{i'i'}M_{jj'} + M_{i'j}M_{i'j'})\right| + \\ & + C\left|(M_{jj}M_{ii'} + M_{ji'}M_{ij})(M_{j'j'}M_{ii'} + M_{j'i'}M_{ij'})\right|. \\ |\mathbb{E}A_{3,ij}A_{3,i'j'}| \leq & C(P_{ii'}P_{jj'} + P_{ij'}P_{i'j})^2.\end{aligned}$$

Given inequality (10) and the symmetry of summation, and statements (a)-(e) proved in Lemma S1.2 in the Supplementary Appendix, we obtain that the last two terms in equation (9) converge to zero. □

Proof of Theorem 3. Denote $\lambda_i = M_i\Pi$, then

$$\widehat{\Phi} = \frac{2}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \left(\eta_i + \Delta\Pi_i\right)\left(M_i\eta + \Delta\lambda_i\right)\left(\eta_j + \Delta\Pi_j\right)\left(M_j\eta + \Delta\lambda_j\right).$$

Let us define $\widehat{\Phi}_0 = \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \eta_i M_i \eta\eta_j M_j\eta$. Assumption 2 guarantees that the variance of $\eta_i = e_i + \Delta\cdot v_i$ is uniformly bounded. Lemma 2 with $\xi_i = (\eta_i, \eta_i, \eta_i)'$ gives $\left|\widehat{\Phi}_0 - \Phi\right| \to^p 0$ uniformly over bounded Δ. Lemma 3 with $\xi_i = (\eta_i, \eta_i, \eta_i, \eta_i)'$ implies $\widehat{\Phi} - \widehat{\Phi}_0 \to^p 0$. □

Lemma 3 *Let $\xi_i = (\xi_i^{(1)}, \xi_i^{(2)}, \xi_i^{(3)}, \xi_i^{(4)})'$ be independent mean zero 4×1 random vectors, such that $\mathbb{E}\|\xi_i\|^4 < C$. Let Assumption 1 hold. Assume that $\lambda'\lambda \leq \frac{C}{K}\Pi'\Pi$ and $\Delta^2 \cdot \frac{\Pi'\Pi}{K} \to 0$ as $N \to \infty$. Then*

$$\begin{aligned}\frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \left(\xi_i^{(1)} + \Delta\Pi_i\right)\left(M_i\xi^{(2)} + \Delta\lambda_i\right)\left(\xi_j^{(3)} + \Delta\Pi_j\right)\left(M_j\xi^{(4)} + \Delta\lambda_j\right) - \\ - \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \xi_i^{(1)} M_i\xi^{(2)}\xi_j^{(3)} M_j\xi^{(4)} \to^p 0.\end{aligned}$$

Proof of Lemma 3. We write the main expression of interest as a polynomial of fourth power in Δ: $\Delta^4 A_4 + \Delta^3 A_3 + \Delta^2 A_2 + \Delta A_1$ and prove that all terms are negligible $\Delta^l A_l \to^p 0$ by showing that their means and variances converge to zero. Notice that for expressions with identical structure but different components of ξ_i, the proof of their negligibility is exactly the same. Thus for simplicity we abuse the notation and drop the superscripts to ξ_i when we can consolidate these expressions. For example, we write the expression for one

of the terms in A_3 as $\frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i \lambda_i \lambda_j \xi_j$, which collects both $\frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i \lambda_i \lambda_j \xi_j^{(1)}$ and $\frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i \lambda_i \lambda_j \xi_j^{(3)}$. We also treat ξ_i in all expressions below as scalar.

$$
\begin{aligned}
A_4 &= \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i \lambda_i \Pi_j \lambda_j; \\
A_3 &= \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i \lambda_i \lambda_j \xi_j + \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i \lambda_i \Pi_j M_j \xi; \\
A_2 &= \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \lambda_i \lambda_j \xi_i \xi_j + \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \lambda_i \xi_i \Pi_j M_j \xi + \\
&\quad + \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \lambda_i \Pi_i \xi_j M_j \xi + \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i \Pi_j M_i \xi M_j \xi; \\
A_1 &= \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \lambda_i \xi_i M_j \xi \xi_j + \frac{1}{K}\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i M_i \xi \xi_j M_j \xi.
\end{aligned}
$$

Term A_4 is deterministic. We use bound (10) and Lemma S1.3 (d):

$$
\Delta^4 |A_4| \leq \frac{C\Delta^4 \Pi'\Pi \lambda'\lambda}{K} \leq \frac{C\Delta^4 (\Pi'\Pi)^2}{K^2} \to 0.
$$

Term A_3 is mean zero. Using the inequality $Var(X+Y) \leq 2Var(X) + 2Var(Y)$ we have:

$$
\begin{aligned}
\Delta^6 Var(A_3) &\leq \frac{C\Delta^6}{K^2} \left(\sum_j \left(\sum_i P_{ij}^2 |\Pi_i||\lambda_i| \right)^2 \lambda_j^2 + \sum_k \left(\sum_i \sum_{j\neq i} \widetilde{P}_{ij}^2 \Pi_i \lambda_i \Pi_j M_{jk} \right)^2 \right) \leq \\
&\leq \frac{C\Delta^6}{K^2} \left((\lambda'\lambda)^2 \Pi'\Pi + \sum_{i,i',j,j'} P_{ij}^2 |\Pi_i \lambda_i \Pi_j| P_{i'j'}^2 |\Pi_{i'} \lambda_{i'} \Pi_{j'}| \sum_k |M_{jk} M_{j'k}| \right) \leq \\
&\leq \frac{C\Delta^6}{K^2} \left((\lambda'\lambda)^2 \Pi'\Pi + (\Pi'\Pi)^2 \lambda'\lambda \right) \leq \frac{C\Delta^6 (\Pi'\Pi)^3}{K^3} \to 0.
\end{aligned}
$$

For the first inequality, we apply Assumption 2 and bound (10). Then we use Cauchy-Schwarz inequality for the first summand: $\left(\sum_i P_{ij}^2 |\Pi_i||\lambda_i|\right)^2 \leq \Pi'\Pi\lambda'\lambda$. For the second summand, we apply Lemma S1.1 (ii) and Lemma S1.3 (c). Finally, we apply Lemma S2.1 and S2.2 to get $\Delta^2 A_2 \to^p 0$ and $\Delta A_1 \to^p 0$. $\square$

Proof of Theorem 4. The infeasible version of AR statistics under $\beta = \beta_0 + \Delta$ is:

$$\begin{aligned}&\frac{1}{\sqrt{K}\sqrt{\Phi}}\sum_i\sum_{j\neq i} P_{ij}e_i(\beta_0)e_j(\beta_0)\\ =&\frac{\Delta^2}{\sqrt{K}\sqrt{\Phi}}\sum_i\sum_{j\neq i} P_{ij}\Pi_i\Pi_j + \frac{2\Delta}{\sqrt{K}\sqrt{\Phi}}\sum_i\left(\sum_{j\neq i} P_{ij}\Pi_j\right)\eta_i + \frac{1}{\sqrt{K}\sqrt{\Phi}}\sum_i\sum_{j\neq i} P_{ij}\eta_i\eta_j. \quad (11)\end{aligned}$$

The first term in (11) is deterministic and equals to $\Delta^2\frac{\mu^2}{\sqrt{K}\sqrt{\Phi}}$. The second term has mean zero and variance

$$\frac{\Delta^2}{K\Phi}\sum_i\left(\sum_{j\neq i} P_{ij}\Pi_j\right)^2 Var(\eta_i) \leq \frac{Cc^2}{K\Phi}\sum_i w_i^2 \leq \frac{C\Pi'\Pi}{K} \to 0.$$

Here we used that variance of η_i is bounded by Assumption 2, $\sum_{j\neq i} P_{ij}\Pi_i = w_i$, and the final bound is proven in Lemma S1.4. Thus, the second term converges to zero in probability uniformly over $|\Delta|^2 \leq c$. The third term in (11) is asymptotically standard normal due to Lemma 1. Finally, we notice that

$$AR(\beta_0) = \sqrt{\frac{\Phi}{\widehat{\Phi}}}\frac{1}{\sqrt{K}\sqrt{\Phi}}\sum_i\sum_{j\neq i} P_{ij}e_i(\beta_0)e_j(\beta_0),$$

and apply Theorem 3. This finishes the proof of statement (4).

Now consider the case when $\frac{\mu^2}{\sqrt{K}\sqrt{\Phi}} \to \infty$ and $\Delta \neq 0$ is fixed. Above we proved that

$$\frac{1}{\sqrt{K}\sqrt{\Phi}}\sum_i\sum_{j\neq i} P_{ij}e_i(\beta_0)e_j(\beta_0) = \frac{\mu^2}{\sqrt{K}\sqrt{\Phi}}\Delta^2 + o_p(1) + O_p(1).$$

Finally, Theorem 3 implies that $\frac{\widehat{\Phi}}{\Phi} \to^p 1$. As a result, we have $AR(\beta_0) \to^p \infty$ when $\frac{\mu^2}{\sqrt{K}\sqrt{\Phi}} \to \infty$ and $\Delta \neq 0$ is fixed. This lead to rejection probability converging to 1. □

Proof of Theorem 5. Denote

$$Q = (Q_{ee}, Q_{Xe}, Q_{XX})' = \frac{1}{\sqrt{K}}\sum_{i=1}^N\sum_{j\neq i} P_{ij}\left(e_ie_j, X_ie_j, X_iX_j\right)'.$$

Lemma A2 in Chao et al. (2012) implies that $\Sigma^{-1/2}\left(Q_{ee}, Q_{Xe}, Q_{XX} - \frac{\mu^2}{\sqrt{K}}\right)' \Rightarrow N(0, I_3)$, where Σ is the asymptotic covariance matrix of Q, with some of its elements written below:

$$\begin{aligned}
\Psi &= \frac{1}{K}\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}^2\gamma_i\gamma_j + \frac{1}{K}\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}^2\sigma_i^2\varsigma_j^2 + \frac{1}{K}\sum_{i=1}^{N}(\sum_{j\neq i} P_{ij}\Pi_j)^2\sigma_i^2 = AVar(Q_{Xe}),\\
\Upsilon &= \frac{2}{K}\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}^2\varsigma_i^2\varsigma_j^2 + \frac{4}{K}\sum_{i=1}^{N}\varsigma_i^2(\sum_{j\neq i} P_{ij}\Pi_j)^2 = AVar(Q_{XX}), \qquad (12)\\
\tau &= \frac{2}{K}\sum_{i=1}^{N}\sum_{j\neq i} P_{ij}^2\varsigma_i^2\gamma_j + \frac{2}{K}\sum_{i=1}^{N}\gamma_i(\sum_{j\neq i} P_{ij}\Pi_j)^2 = ACov(Q_{Xe}, Q_{XX}), \quad \varrho = \frac{\tau}{\sqrt{\Psi}\sqrt{\Upsilon}}.
\end{aligned}$$

Note that $\widehat{e}_i = Y_i - X_i\widehat{\beta}_{JIVE} = e_i - X_i(\widehat{\beta}_{JIVE} - \beta)$ and $(\widehat{\beta}_{JIVE} - \beta_0) = Q_{Xe}/Q_{XX}$. Thus,

$$Wald(\beta_0) = \frac{Q_{Xe}^2}{\sum_{i=1}^{N}\left(\sum_{j\neq i} P_{ij}X_j\right)^2\frac{\widehat{e}_iM_i\widehat{e}}{M_{ii}} + \sum_{i=1}^{N}\sum_{j\neq i}\widetilde{P}_{ij}^2M_iX\widehat{e}_iM_jX\widehat{e}_j},$$

where the denominator expands to

$$\begin{aligned}
&\sum_{i=1}^{N}\left(\sum_{j\neq i} P_{ij}X_j\right)^2\frac{\widehat{e}_iM_i\widehat{e}}{M_{ii}} + \sum_{i=1}^{N}\sum_{j\neq i}\widetilde{P}_{ij}^2M_iX\widehat{e}_iM_jX\widehat{e}_j =\\
&= \left\{\sum_{i=1}^{N}\left(\sum_{j\neq i} P_{ij}X_j\right)^2\frac{e_iM_ie}{M_{ii}} + \sum_{i=1}^{N}\sum_{j\neq i}\widetilde{P}_{ij}^2M_iXe_iM_jXe_j\right\} -\\
&-\frac{Q_{Xe}}{Q_{XX}}\left\{\sum_{i=1}^{N}\left(\sum_{j\neq i} P_{ij}X_j\right)^2\left(\frac{e_iM_iX}{M_{ii}} + \frac{X_iM_ie}{M_{ii}}\right) + 2\sum_{i=1}^{N}\sum_{j\neq i}\widetilde{P}_{ij}^2M_iXe_iM_jXX_j\right\} +\\
&+\frac{Q_{Xe}^2}{Q_{XX}^2}\left\{\sum_{i=1}^{N}\left(\sum_{j\neq i} P_{ij}X_j\right)^2\frac{X_iM_iX}{M_{ii}} + \sum_{i=1}^{N}\sum_{j\neq i}\widetilde{P}_{ij}^2M_iXX_iM_jXX_j\right\}.
\end{aligned}$$

Applying Lemma S3.1 from the Supplementary Appendix to the expanded expression of the denominator, we show the terms appearing in the braces converge to $\Psi, 2\tau$ and Υ respectively. Then

$$Wald(\beta_0) = \frac{Q_{Xe}^2}{\Psi - 2\frac{Q_{Xe}}{Q_{XX}}\tau + \frac{Q_{Xe}^2}{Q_{XX}^2}\Upsilon}(1 + o_p(1)) = \frac{Q_{Xe}^2/\Psi}{1 - 2\frac{Q_{Xe}/\sqrt{\Psi}}{Q_{XX}/\sqrt{\Upsilon}}\varrho + \frac{Q_{Xe}^2}{Q_{XX}^2}\frac{\Upsilon}{\Psi}}(1 + o_p(1)).$$

Lemmas 2 and 3 applied to $\widehat{\Upsilon}$ with $\xi_i = (v_i, v_i, v_i, v_i)'$ and $\Delta = 1$ give $\widetilde{F} = \frac{Q_{XX}}{\sqrt{\Upsilon}}(1 + o_p(1))$. Thus, the statement of Theorem 5 holds where we denote $\left(\xi, \nu - \frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}}\right)$ to be the Gaussian limit of $(\frac{Q_{Xe}}{\sqrt{\Psi}}, \frac{Q_{XX}}{\sqrt{\Upsilon}} - \frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}})$. □

Proof of Corollary 1. Denote $x = \frac{\mu^2}{\sqrt{K}\sqrt{\Upsilon}}$. If $x > 2.5$ then due to Theorem 5:

$$\mathbb{P}_x\{\widetilde{F} > 4.14 \text{ and } Wald(\beta_0) \geq \chi^2_{1,0.95}\} \leq \mathbb{P}_x\{Wald(\beta_0) \geq \chi^2_{1,0.95}\} \leq 0.10.$$

If $x \leq 2.5$ then due to the asymptotic gaussianity of $\widetilde{F}$:

$$\mathbb{P}_x\{\widetilde{F} > 4.14 \text{ and } Wald(\beta_0) \geq \chi^2_{1,0.95}\} \leq \mathbb{P}_x\{\widetilde{F} > 4.14\} \leq 0.05.$$

Finally, for any $x > 0$:

$$\mathbb{P}_x\left\{H_0 \text{ is rejected }\right\} = \mathbb{P}\{\widetilde{F} > 4.14 \text{ and } Wald(\beta_0) \geq \chi^2_{1,0.95}\} +$$
$$+\mathbb{P}\{\widetilde{F} > 4.14 \text{ and } AR(\beta_0) \leq z^2_{1,0.95}\} \leq 0.10 + \mathbb{P}\{AR(\beta_0) \leq z^2_{1,0.95}\} \leq 0.15.$$

8.1 Simulation Details

To create many instruments, we interact QOB dummies with dummies for year of birth (YOB) and place (state) of birth (POB). Interacting three QOB dummies with nine YOB and 50 POB dummies generates 180 excluded instruments. The excluded instruments are

$$Z_i = ((\mathbf{1}\{Q_i = q, C_i = c\})'_{q\in\{2,3,4\}, c\in\{31,\ldots,39\}}, \mathbf{1}\{Q_i = q, P_i = p\})'_{q\in\{2,3,4\}, p\in\{50 \text{ states}\}})',$$

where Q_i, C_i, P_i are i's QOB, YOB and POB respectively. Note, that Z_i are not group instruments in the strict sense as they are not mutually exclusive. We exclude instruments with $\sum_{i=1}^N Z_{ij} < 5$ to satisfy the balanced instruments assumption (Assumption 1).

To increase the amount of omitted variable bias, we follow Angrist and Frandsen (2019) by taking the LIML model as the ground truth, where the outcome variable is Y_i (income), the endogenous variable X_i (highest grade completed) is instrumented by Z_i and the control variables are a full set of POB-by-YOB interactions. Specifically, starting

with the full 1980 census sample, we compute the average X_i in each QOB-YOB-POB cell $\bar{s}(q,c,p)$. We then estimate LIML and retain $\hat{y}(c,p)$, the second-stage fitted value after subtracting $\hat{\beta}_{LIML}X_i$ where $\hat{\beta}_{LIML}$ is the LIML estimate of the returns to schooling. We also retain the variance of LIML residuals $\omega(Q_i,C_i,P_i)$ to mimic the heteroskedasticity.

The simulation model we consider is then

$$\tilde{y}_i = \bar{y} + 0.1\tilde{s}_i + \omega(Q_i,C_i,P_i)(\nu_i + \kappa_2\epsilon_i)$$
$$\tilde{s}_i \sim Poisson(\mu_i),$$

for independent standard normal ν_i and ϵ_i. Here $\bar{y} = \frac{1}{N}\sum_i \hat{y}(C_i,P_i)$ and $\mu_i = \max\{1, \gamma_0 + \gamma_Z'Z_i + \kappa_1\nu_i\}$ where $\gamma_0 + \gamma_Z'Z_i$ is the projection of $\bar{s}(Q_i,C_i,P_i)$ onto a constant and Z_i. We set $\kappa_1 = 1.7$ and $\kappa_2 = 0.1$ following Angrist and Frandsen (2019).