Instrument Strength in IV Estimation and Inference: A Guide to Theory and Practice

Michael Keane[†] and Timothy Neal[†]

[†] *CEPAR & School of Economics, University of New South Wales*
E-mail: m.keane@unsw.edu.au

November 30, 2022

Abstract 2SLS has poor properties if instruments are exogenous but weak. But how strong must instruments be for 2SLS estimates and test statistics to exhibit acceptable properties? A common standard is a first-stage $F \geq 10$. This is adequate to ensure two-tailed t-tests have small size distortions. But other problems persist: In particular, we show 2SLS standard errors tend to be artificially small in samples where the estimate is most contaminated by the OLS bias. Hence, if the bias is positive, the t-test has little power to detect true negative effects, and inflated power to find positive effects. This phenomenon, which we call a "power asymmetry," persists even if first-stage F is in the thousands. Robust tests like Anderson-Rubin perform better, and should be used in lieu of the t-test even with strong instruments. We also show how 2SLS test statistics typically suffer from very low power when first-stage F is near 10, leading us to suggest a higher standard of instrument strength in empirical practice.

JEL: *C12, C26, C36*

Keywords: *Instrumental variables, weak instruments, 2SLS, endogeneity, F-test, size distortion, Anderson-Rubin test, likelihood ratio test, LIML, GMM, Fuller, JIVE*

1. INTRODUCTION

Economists frequently use instrumental variable (IV) methods to deal with endogeneity problems, and the most popular method is two stage least squares (2SLS). But the work of Bound et al. (1995) and Nelson and Startz (1990) highlights the poor properties of 2SLS when instruments are exogenous but "weak," in the sense of being weakly correlated with the endogenous variable. Two problems have received tremendous attention: size inflation in 2SLS t-tests and median bias of 2SLS estimates towards OLS. A large literature has emerged on (i) testing if instruments are strong enough to avoid these problems, and (ii) developing statistical tests that are robust to weak IV problems.

We provide an accessible guide to this literature. Our main contribution is to highlight key problems with 2SLS t-tests that weak IV tests gloss over. We show how 2SLS standard errors tend to be artificially small in samples where the estimate is most contaminated by the OLS bias. Thus, if the OLS bias is positive, the t-test has inflated power to find false positive effects, and poor power to detect true negative effects, even in large samples where instruments are "strong" by conventional standards. Fortunately, robust tests like Anderson and Rubin (1949) can avoid this problem and are easy to implement. We argue they should be widely adopted in lieu of the t-test – even if instruments are strong.

In an important paper, Staiger and Stock (1997) showed that t-tests suffer from severe size distortions if instruments attain only 5% significance in the 2SLS first stage (i.e, a first stage F of 3.84 in the single instrument case). This led them to advocate a higher standard of instrument relevance. They find if first-stage F is 10 then a 2SLS 5% two-tailed t-test rejects a true null H_0:$\beta = 0$ at a rate not "too far" from the correct 5% rate. Thus, Stock and Watson (2015, p.490) write: "One simple rule of thumb is that you do need not to worry about weak instruments if the first stage F-statistic exceeds 10."

Later, Stock and Yogo (2005) derived thresholds for first-stage sample $\hat{F}$ based on the maximal size distortion in t-tests one is willing to tolerate (i.e., How often does a 5% t-test test reject?). For example, in the single instrument case they show that $\hat{F} > 16.4$ ensures (with 95% confidence) that a two-tailed 5% t-test rejects a true null H_0:β=0 at a rate no higher than 10%. Recently, Lee et al. (2022) show a much higher standard of $\hat{F} > 104.7$ is required to insure t-tests reject at a rate no higher than the correct 5% rate.

Stock and Yogo (2005) weak instrument tests are designed to assess size distortions in two-tailed 2SLS t-tests.[1] The weak IV literature in general has focused heavily on size inflation of t-tests and bias in 2SLS estimates. We argue this focus on size and bias has caused the literature to gloss over *other* problematic properties of 2SLS that persist even when instruments are "strong" according to Stock-Yogo tests, and even in large samples.

In particular, 2SLS suffers from two key problems if first-stage $\hat{F}$ is in the 10 to 20 range typically deemed acceptable. First, estimates are imprecise. Second, a strong association exists between 2SLS estimates and standard errors: *2SLS generates artificially low standard errors in samples where the estimate is most contaminated by endogeneity.*

The strong association between 2SLS estimates and standard errors that we identify persists even if instruments are very strong. It has two important consequences: 2SLS estimates shifted towards OLS will appear spuriously precise, so the t-test has inflated power to judge such estimates significant. Conversely, 2SLS t-tests have little power to detect a true β opposite in sign to the OLS bias. This phenomenon, which we call a "power asymmetry," renders the t-test unreliable even if instruments are quite strong.[2]

The 2SLS power asymmetry has serious implications: In an archetypal application of IV, one seeks to test if a program has a positive effect on an outcome, but a confound arises because those who participate are positively selected on unobservables. Even if instruments are quite strong by conventional standards, the 2SLS t-test will have inflated power to find false positive effects, and little power to detect true negative effects.

In the single instrument (exactly identified) case we show that the weak instrument robust test of Anderson and Rubin (1949) greatly alleviates the power asymmetry problem that plagues the t-test, making it far more reliable. To illustrate, we also provide an empirical application to estimating the effect of anticipated income changes on consumption. This clearly shows the superiority of the AR test over the t-test – it not only has correct size but also substantially better power properties. Furthermore, the AR test is very easy to calculate. Thus we argue the AR test should be widely adopted in lieu of the t-test in the exactly identified case, even if instruments are strong.

Finally, we consider the over-identified case. The use of multiple instruments increases efficiency. But it also increases the bias of 2SLS towards OLS, and the size distortion and power asymmetry in t-tests. The limited information maximum likelihood (LIML) estimator of Anderson and Rubin (1949), in conjunction with the conditional likelihood ratio (CLR) test of Moreira (2003), give much more reliable results. Contrary to widespread misperception among applied researchers, we show that both LIML and CLR are simple to interpret and implement.[3] We argue these methods should be widely adopted, not only when instruments are weak but also when instruments are strong.

[1] They also assess bias of 2SLS in models with overidentification of degree 2, so the mean of 2SLS exists.

[2] As Angrist and Kolesár (2021) note, our concept of power asymmetry is non-standard, as it requires one to consider the behavior of test statistics *conditional* on the parameter estimate. Nevertheless, we argue this power asymmetry has very important implications for empirical work.

[3] Finlay and Magnusson (2009) provide Stata code for a heteroskedasticy robust version of CLR. The continuously updated GMM of Hansen et al. (1996) generalizes LIML to heteroskedastic data.

2. BACKGROUND: PROBLEMATIC PROPERTIES OF 2SLS T-TESTS

Consider an exactly identified linear IV model, where endogenous variable y is regressed on a single endogenous variable x, and there is a single exogenous instrument z. We focus on this simple case as it clarifies the key ideas, and it is the most common in applied practice. The parameter $\rho \in [0,1]$ controls the extent of the endogeneity problem, while π governs the relationship between the instrument and the regressor:

$$\begin{aligned} y_i &= \beta x_i + u_i \\ x_i &= \pi z_i + e_i \quad \text{where} \quad e_i = \rho u_i + \sqrt{1-\rho^2}\eta_i \\ u_i &\sim iid\, N(0,1), \eta_i \sim iid\, N(0,1), z_i \sim iid\, N(0,1) \end{aligned} \tag{1}$$

We assume *iid*-normal errors to allow analytic power calculations. We normalize all error variances to 1.0 as this allows us to interpret β as roughly the standard deviation change in y induced by a one sigma change in x. In much of our analysis it will be useful for the magnitude of β to be interpretable. The OLS bias is $E(\hat{\beta}_{OLS} - \beta) = \rho$.

This *iid* normal setup is not as restrictive as it may appear, as Andrews et al. (2019) show that for any heteroskedastic DGP, there exists a homoskedastic DGP yielding equivalent behavior of 2SLS estimates and test statistics. Furthermore, any exogenous covariates can be partialed out of y and x without changing anything of substance.

The 2SLS estimator of β takes the following form, where $\hat{}$ denotes a sample value:

$$\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i x_i} = \beta + \frac{\sum_{i=1}^n z_i u_i}{\sum_{i=1}^n z_i x_i} = \beta + \frac{\widehat{cov}\,(z,u)}{\widehat{cov}\,(z,x)} \tag{2}$$

This may be obtained via a two-step process where in the first-stage one regresses x on z, and in the second stage one regresses y on the fitted values from the first stage.

The F statistic from the population first-stage regression of x on z determines the strength of the instrument. It is often called the concentration parameter, C. Of course, we cannot observe F, as we can only observe the sample $\hat{F}$ from the first-stage regression. $\hat{F}$ provides an estimate of the strength of the instrument. In the notation of (1) we have:

$$C = F = N\frac{Var(z\pi)}{\sigma_e^2} = N\frac{\pi^2\sigma_z^2}{\sigma_e^2} = N\pi^2 \text{ and } \hat{F} = N\hat{\pi}^2\hat{\sigma}_z^2/\hat{\sigma}_e^2$$

The size of a statistical test is the probability of rejecting a true null hypothesis. Unfortunately, the size of a 5% level 2SLS t-test is not generally 5%. The 2SLS t-test is not a pivotal statistic, so its size depends on the nuisance parameters ρ and C.[4]

Stock and Yogo (2005) studied how the size of the 5% two-tailed 2SLS t-test deviates from the correct 5% level, and how this size distortion depends on instrument strength. They derive a formula for power of the t-test in terms of C, ρ and true β that we present in Appendix A, equation (A2). Evaluating power at β=0 gives the size of the test.

A complication arises because as size is increasing in $|\rho|$, so Stock-Yogo focus on the *maximal* size distortion, which occurs when $\rho = \pm 1$. The integral in (A2) can then be evaluated numerically to determine how the size of the t-test depends on C. For example, doing a grid search over F to set size approximately equal to 15%, they obtain F=1.82. So this level of instrument strength guarantees a maximal size distortion of 10%.

Stock and Yogo (2005) weak IV tests are sample $\hat{F}$ thresholds that give 95% confidence the population F is above some threshold that, in turn, implies a maximal size distortion.

[4]The 2SLS t-test is only asymptotically pivotal as C grows large. In contrast, the OLS t-test is pivotal, as its distribution is purely a function of the data.

The sample $\hat{F}$ is a draw from a non-central F with non-centrality parameter C. Thus, for example, $\hat{F} > 8.96$ gives 95% confidence that F is at least 1.82, which in turn, implies the maximum size of a two-tailed 5% t-test is 15% (i.e., maximal size distortion of 10%).

Table 1 give examples of different Stock-Yogo test thresholds and the maximal size that each achieves. For example, $\hat{F} > 10$, which corresponds to the popular Staiger-Stock rule of thumb for acceptable instrument strength, gives 95% confidence that that C is at least 2.3, which in turn, insures a maximal size of 13.5% (distortion of 8.5%).

Table 1. Stock-Yogo Test: Example First-Stage $\hat{F}$ Critical Values

C = Pop F	π	$\hat{F}_{.05}$ critical value	Max Size
1.82	0.0427	8.96	15%
2.30	0.0480	10.00	13.5%
5.78	0.0760	16.38	10%
10.00	0.1000	23.10	8.6%
29.44	0.1716	50.00	6.4%
73.75	0.2716	104.70	5%

Note: Instrument is significant at the 5% level if first-stage $\hat{F} > 3.84$. Higher levels of $\hat{F}$ reduce maximal size distortion to levels in column 4. The π levels are specific to N=1000, as $C = N\pi^2$.

The limitation of the Stock-Yogo analysis is it emphasizes the level of the power curve at the single point β=0 and $\rho = \pm 1$. Our goal is to obtain a broader view of the properties of the t-test. We begin with a simple experiment: We simulate 10,000 artificial data sets of size N=1000 from the model in (1), assuming β=0, and setting the degree of endogeneity ρ=0.80 so the OLS bias is positive. We set π=0.048 so $C = F = N\pi^2 = 2.3$. This level of C is interesting, as the popular $\hat{F} > 10$ rule is actually a 5% level test for $F \geq 2.3$. We then run 2SLS on each artificial data set and summarize the results.

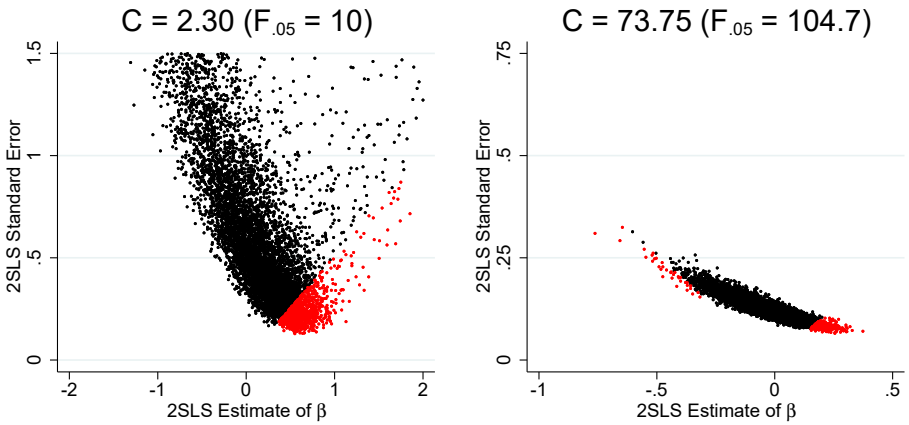
The left panel of Figure 1 plots the 2SLS standard error estimates $se(\hat{\beta}_{2SLS})$ against the 2SLS estimates $\hat{\beta}_{2SLS}$. This reveals a striking pattern: A strong negative association is evident; in fact, the Spearman r_s is -0.576 and Kendall's τ is -0.511. Notice the magnitude of the variation in the standard errors is substantial, as the y-axis shows standard errors that range from about 0.2 to 1.5. Thus, 2SLS estimates that are most shifted toward the OLS bias appear to be *much* more precisely estimated.

The red dots in Figure 1 indicate runs where $\hat{\beta}_{2SLS}$ differs significantly from zero according to a two-tailed 5% t-test. The null hypothesis H_0:$\beta = 0$ is rejected at a 10% rate, so as expected there is a modest size inflation. But this masks a deeper problem: Due to the negative association between the 2SLS estimates and their standard errors, <u>all</u> rejections occur when $\hat{\beta}_{2SLS}$>0, and <u>none</u> when $\hat{\beta}_{2SLS}$<0. Only the estimates most shifted towards the OLS bias are ever judged significant.

The association between 2SLS estimates and standard errors is not a weak instrument phenomenon. The right panel of Figure 1 plots results for the very strong instrument case of C=74 ($F_{.05}$=105). A strong negative association persists. In fact, Spearman's r_s is -0.92 and Kendall's τ is -0.75. The 2SLS t-test now has a rejection rate of 4.87%, so size is roughly correct. But 93% of those rejections occur when $\hat{\beta}_{2SLS} > 0$. In Keane and Neal (2022b) we show this association persists even with C in the thousands.

The source of the negative association between 2SLS estimates and standard errors is simple to understand. In our DGP $cov(z, x) > 0$. In our simulation $\widehat{cov}(z, x) > 0$ in all runs with C=74, and 93.3% of runs with C=2.3. Provided $\widehat{cov}(z, x) > 0$, equation (2)

Figure 1. Standard Error of $\hat{\beta}_{2SLS}$ plotted against $\hat{\beta}_{2SLS}$ itself ($\rho = 0.80$)



Note: Runs with standard error > 1.5 *not shown. Red dots indicate* $H_0 : \beta = 0$ *rejected at 5% level.*

shows how a positive realization of $\widehat{cov}(z, u)$, the sample covariance between the instrument and the structural error, generates an estimate shifted in the positive direction. This is the direction of the OLS bias, as $E(\hat{\beta}_{OLS}) = \rho = 0.80$.

Crucially, a positive sample realization of $\widehat{cov}(z, u)$ also drives up the sample covariance between the instrument and the endogenous variable. This makes the instrument appear spuriously strong, as is obvious because:

$$\widehat{cov}\,(z, x) = \pi\widehat{var}\,(z) + \rho\widehat{cov}\,(z, u) + \sqrt{1 - \rho^2}\widehat{cov}\,(z, \eta) \tag{3}$$

This spurious instrument strength drives down the 2SLS standard error.[5,6]

Thus, a positive sample realization of $\widehat{cov}(z, u)$ generates both (i) an estimate shifted towards OLS and (ii) a low estimated standard error. Hence, 2SLS will appear "spuriously precise" in samples where the estimate is most shifted in the direction of the OLS bias.[7] This generates the negative association that we see in Figure 1.

We now run a second experiment that illustrates the power asymmetry problem that afflicts 2SLS as a consequence of the association between the estimates and their standard errors. We consider DGPs where the true β is ± 0.30, and assess the power of the 2SLS t-test to reject the false null H_0:β=0. Importantly, these values of β would be quantitatively large, but plausible, in typical empirical applications, as they imply a one standard deviation change in x induces an 0.25 standard deviation change in y. We consider all 5 levels of instrument strength in Table 1, and three levels of endogeneity, $\rho \in (0, 0.5, 1.0)$. For each parameter setting we simulate 10,000 artificial data sets with N=1000.

We summarize the results in Table 2. A striking result is that a 2SLS t-test has almost no power to detect a sizeable true negative effect when the OLS bias is positive, unless instrument strength is far above conventional weak IV test thresholds. For example, if C=10 and ρ=0.5, the probability of rejecting H_0:β=0 is 23.7% when $\beta = 0.30$ compared

[5] As $Var(\hat{\beta}_{2SLS}) = Var(\hat{\beta}_{OLS})/R^2_{z,x}$, the larger is $\widehat{cov}(z, x)$ the smaller is the standard error.

[6] If C=2.3 the chance of $\widehat{cov}(z, x) < 0$ is 6.7%. These runs typically generate large positive estimates with large standard errors. But this occurs too infrequently to change the overall negative association.

[7] We call this phenomenon "spurious precision" as a positive $\widehat{cov}(z, u)$ makes the instrument appear stronger than it really is – i.e., it drives up $\widehat{cov}(z, x)$ relative to $cov(z, x)$. This drives down the standard error, so it understates the actual degree of uncertainty about the parameter estimate.

Table 2. Power of 2SLS t-Test – Frequency of Rejecting H_0:$\beta = 0$ (%)

Concentration Parameter	$F_{5\%}$	$\beta = 0.3$			$\beta = -0.3$		
		$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$
1.82	8.96	1.8	11.7	25.5	1.7	0.1	4.2
2.30	10.00	2.4	13.0	25.1	2.2	0.2	3.2
5.78	16.38	7.2	18.8	26.3	7.2	0.5	0.8
10.00	23.10	13.4	23.7	28.9	13.3	2.3	0.2
73.75	104.7	71.4	67.8	65.1	71.9	78.0	89.1

Note: The table reports the probability of rejecting the false null hypothesis H_0*:* $\beta = 0$.

to only 2.3% when β=-0.30. This power asymmetry arises from the geometry of Figure 1. If β=0.30 the cloud of points shifts right, while if β=-0.30 it shifts left. Clearly, a rightward shift generates more significant results.

In summary, the 2SLS estimator has the unfortunate property that it tends to generate standard errors that are too low precisely when it also generates estimates shifted in the direction of the OLS bias. As a consequence, it is difficult for a 2SLS t-test to detect plausibly sized true negative effects when the OLS bias is positive. This pattern is reversed if the OLS bias is negative. We explore this problem further in the next section.

3. THE ANDERSON-RUBIN TEST VS. THE T-TEST

The usual suggestion of the theory literature is to avoid the t-test if instruments are weak, and instead use robust tests that have correct size regardless of instrument strength. In the single instrument case the unambiguous choice is the Anderson and Rubin (1949) test. The AR test is based on the reduced-form regression of y on z, which is $y = z\beta\pi + (\beta e + u) = z\xi + v$ where $\xi = \beta\pi$. Given a valid instrument z, which must satisfy $\pi \neq 0$ and $cov(z, v) = 0$, a test of the null hypothesis H_0:$\xi = 0$ provides an alternative way to test H_0:$\beta = 0$. Thus, The AR test judges $\hat{\beta}_{2SLS}$ to be significant if z is a significant predictor of y in the reduced form. Equivalently, the AR test is simply the F-test from the regression of y on $z\hat{\pi}$, where $\hat{\pi}$ is the first stage estimate of π.

The AR test has correct size, regardless of instrument strength, as it is simply an F-test from OLS regression of y on z. It is a pivotal statistic, meaning it does not depend on C or ρ. The AR test also has superior power properties relative to the t-test, as illustrated in Figure 2. It presents analytical power curves for both tests, for the model in (1), obtained as described in Appendix A. We set the level of instrument strength to C=10, which is well above conventional weak IV thresholds. The left and right panels show results for $\rho = 0.50$ and 0.80, corresponding to moderate and severe endogeneity problems, respectively. We adopt a 5% level for both tests.

An unbiased statistical test has the desirable property that the probability of rejecting H_0:$\beta = 0$ is minimized if the true β is in fact zero. We can see in Figure 2 that the AR test is unbiased. It also has correct size, as its power evaluated at $\beta = 0$ is exactly 5%.

In contrast, the t-test is biased. As we see in Figure 2, left panel, if $\rho = 0.50$ the power of the t-test is near zero when the true β is in the vicinity of -0.25. So the probability of rejecting H_0:$\beta = 0$ is minimized when true β is near -0.25 rather than at zero. And if $\rho = 0.80$ (right panel) t-test power is near zero for true β in the -0.25 to -0.40 range. Recall that β is roughly the standard deviation change in y induced by a one sigma change x. Effect sizes of -0.25 to -0.40 are quite large in typical applications. Thus,

these results show the t-test has almost no power to detect a wide range of substantively large true negative effects when the endogeneity bias afflicting OLS is positive. This is due to the negative association between 2SLS estimates and standard errors.

In contrast, as we also see in Figure 2, the AR test has far better power than the t-test to detect true negative effects. In the single instrument case Moreira (2009) shows AR is the uniformly most powerful unbiased test (for testing a point null hypothesis). It has better power than any other unbiased test, regardless of the true parameter value.

Figure 2. Power of t-Test vs. AR-Test when $C = 10$ ($F_{5\%}$=23.1)

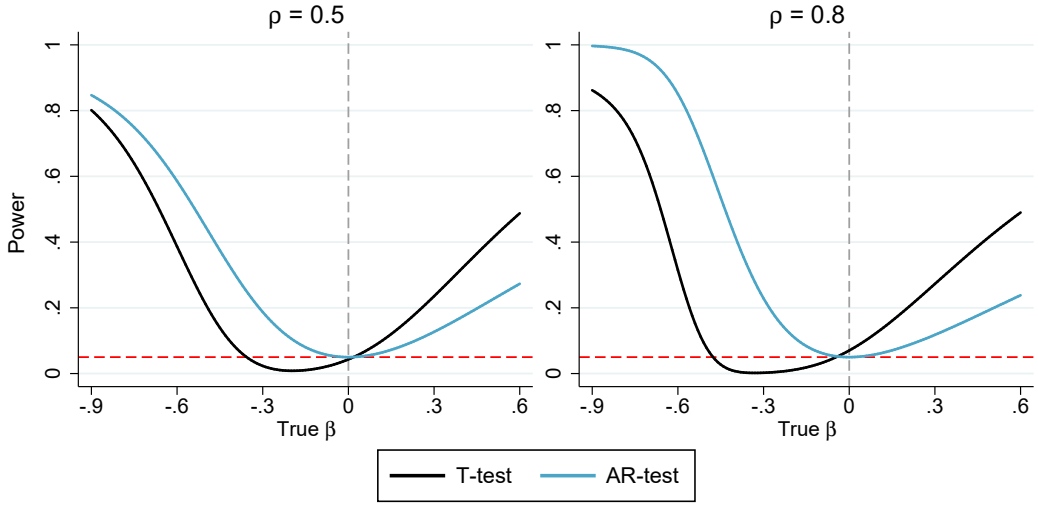


Figure 2 also shows that if true β is positive the t-test has higher power than AR against H_0:β=0. We argue this property is not desirable, as it reflects that 2SLS standard errors are spuriously small for estimates shifted in the direction of the OLS bias (positive).

To clarify this point, consider an unidentified model. Figure 3 presents results for C=0, so the instrument z is independent of the endogenous variables x and y. The power of the AR test is, appropriately, a flat line at 5% independent of the level of β. The chance of concluding $\hat{\beta}$ is significant is exactly the probability of a significant covariance between y and z arising by chance. In contrast, the t-test rejects H_0:$\beta = 0$ at a 13.2% rate if β=0, and its power is strongly increasing in β, rising to 39.4% if $\beta = 1$.

What generates t-test power in an unidentified model? The right panel shows results of 10,000 simulation runs for the specific case of $\beta = 1$. Here $E(\hat{\beta}_{OLS})$=1.8. Remarkably, the $\hat{\beta}_{2SLS}$ in the near vicinity of 1.8 appear to be rather precisely estimated, with a median standard error of roughly 0.49, despite the fact the model is not identified.[8] This is what we refer to as "spurious precision." The significant estimates are shaded red (39.4%). A striking fact is that only estimates *larger* than the true value are judged significant by the t-test. Thus, large realizations of $\widehat{cov}(z,u)$ generate upward biased estimates that seem precise because $\widehat{cov}(z,x)$>0 is spuriously high. The t-test often calls these estimates significant, which we call "spurious power." This pattern persists in identified models:[9]

[8] As a point of comparison, 90% of estimates fall in the interval 1.8 ±3.7, and 95% fall in 1.8 ±7.4. So a standard error of 0.49 greatly understates the level of uncertainty about the parameter estimate.

[9] Phillips (1989) notes behavior of 2SLS in the unidentified case impacts its behavior in identified cases.

2SLS standard errors are spuriously small when the estimate is shifted in the direction of the OLS bias, which gives the t-test spuriously high power in that direction.

Figure 3. Power of the T-Test vs. AR-Test when Model is Not Identified ($C = 0$)

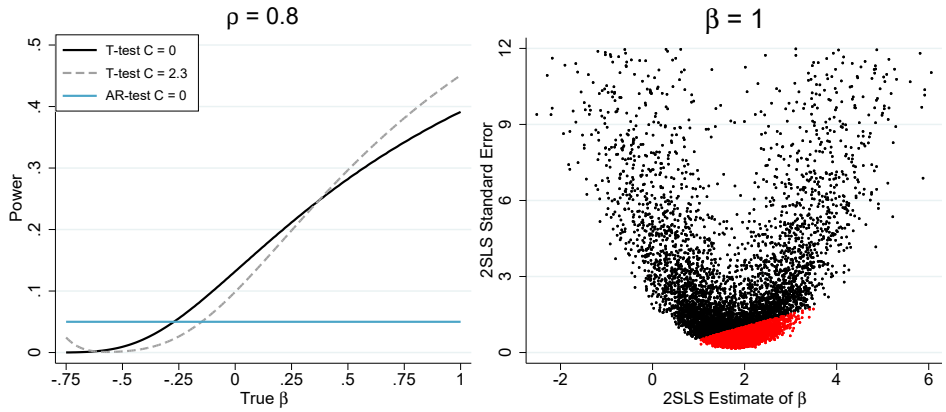


Figure 3 also plots a t-test power curve for C=2.3 (dashed line). Remarkably, the t-test power curves for C=0 and C=2.3 look similar over a wide range of β. Notably, if C=2.3 there is a 33% chance first-stage $\hat{F}$ will exceed the 5% critical value 3.84, giving 95% confidence our model is identified. Logically, then, a 5% t-test should not indicate $\hat{\beta}$ is significant more than 33% of the time. But t-test power clearly exceeds this logical upper bound if β is large enough. The AR test does not have this problem.

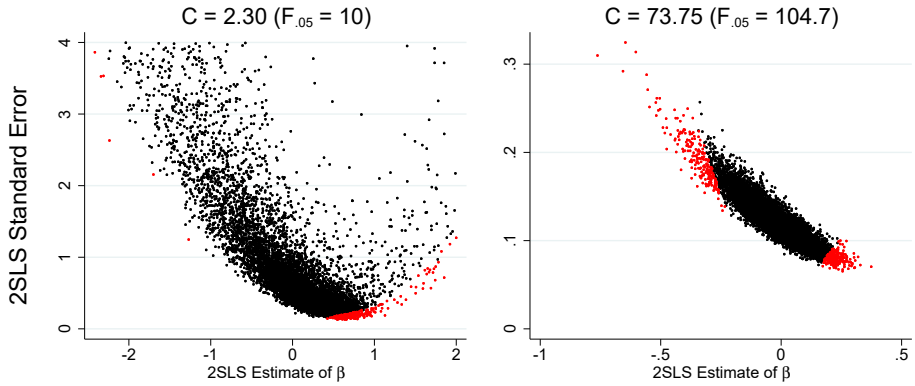
The Staiger-Stock-Yogo approach avoids this problem by ignoring t-test results unless first-stage $\hat{F}$ exceeds a threshold – see Table 1. Otherwise, we do not reject the null. Then, applying even a minimal $\hat{F}$>3.84 threshold, t-test power would never exceed 5% when C=0 or 33% when C=2.3. However, $\hat{F}$ tends to be larger when $\hat{\beta}_{2SLS}$ is near $E(\hat{\beta}_{OLS})$, creating a pre-test bias. So a better and simpler solution is to adopt the AR test.

Figure 2 also shows that power of the two-tailed 5% t-test evaluated at $\beta = 0$ is close to 5% for both ρ=0.5 and ρ=0.8. This illustrates Angrist and Kolesár (2021)'s point that two-tailed t-test size inflation is minor unless the instrument is weak and endogeneity very severe. If ones' only concern is that the power curve is not too far above 5% at the single point β=0, then one might argue, as they do, that the t-test is not too bad. But as we have seen, inspection of the whole power curve reveals deeper problems.

Finally, we return to the issue of balance between positive vs. negative rejections. Figure 2 conceals this, as the power of a two-tailed test is defined as their sum. Recall that in Figure 1 we found that almost all t-test rejections of H$_0$:β=0 occur when $\hat{\beta}_{2SLS} > 0$, reflecting the severe power asymmetry of the t-test. Figure 4 is identical, except now the red shaded area indicates cases where the AR test rejects the null. The left panel reports the case of ρ=0.80 and C=2.3. As expected, the AR test rejects H$_0$:β=0 at the correct 5% rate. But 85% of those rejections occur when $\hat{\beta}_{2SLS} > 0$. So, surprisingly, the AR test seems to suffer from the same power asymmetry problem as the t-test.

The reason for the power asymmetry in the AR test is again the strong positive association between $\rho\widehat{cov}(z,u)$ and $\hat{\beta}_{2SLS}$. A large value of $\rho\widehat{cov}(z,u)$ tends to generate a large value of the AR statistic, as it tends to make z appear more significant in the reduced form for y. Thus, if ρ >0 the AR test and $\hat{\beta}_{2SLS}$ have a positive association, making the AR test more likely to reject H$_0$:β=0 if $\hat{\beta}_{2SLS} > 0$.

Figure 4. AR-test rejections: $SE(\hat{\beta}_{2SLS})$ plotted against $\hat{\beta}_{2SLS}$ ($\rho = 0.80$)



Note: Runs with standard error > 4 are not shown. Red dots indicate $H_0 : \beta = 0$ is rejected at the 5% level. Results are for the 2SLS t-test (left panel) or the Anderson-Rubin test (right panel).

In contrast the t-test, AR test power asymmetry vanishes quickly as instrument strength increases. The right panel of Figure 4 shows results for the strong instrument case of C=74. Here the AR test exhibits a fairly even balance of positive (54%) vs. negative rejections. So AR achieves this balance at a vastly smaller first-stage F than the t-test.

To summarize, we have seen that the AR test exhibits far superior power properties to the t-test, not only when instruments are weak but also when instruments are strong. Not only does the AR test have correct size, but it also has far better power to detect true effects opposite in sign to the OLS bias. Based on these considerations, we recommend that applied reasearchers should adopt the AR in lieu of the t-test even when the first-stage F-statistic is far above conventional standards for a strong instrument.

We conclude this section with some important observations on the AR test. First, we note that Moreira (2009)'s optimal power result applies to *iid* settings. However, Moreira and Moreira (2019) extend it to settings with heteroskedasticity and clustering. In that case, one should implement the AR test using a heteroskedasticity and/or cluster robust F-test, as we illustrate in Section 4 and in Keane and Neal (2022a).

Second, the AR statistic is pivotal, as its distribution under H_0:β=0 does not depend on ρ and C. This allows one to invert the AR test to form valid confidence intervals, as discussed in Anderson and Rubin (1949) and Dufour (2004). We illustrate the use of AR confidence intervals in Section 4. In contrast, the distribution of the t-statistic is highly dependent on ρ and C, rendering confidence intervals suspect even in large finite samples with weak (or moderately strong) instruments.

If instruments are very weak the AR confidence interval can be unbounded. In particular, if the first-stage $\hat{F}$ <3.84 then a 95% confidence interval for β is unbounded. As Dufour (2004) notes, this is not a problem but rather an accurate reflection of uncertainty. If $\hat{F}$ <3.84 we do not have 95% confidence that the instrument is significant in the first stage, so we lack 95% confidence that the model is identified. It is an odd property of the t-test that it gives a bounded confidence interval in this case.

Third, with multiple instruments the AR test is no longer optimal. Instead, Moreira (2003) shows the conditional likelihood ratio (CLR) test has correct size and better power than AR. These tests are equivalent in the single instrument case. We compare the performance of t, AR and CLR tests in the over-identified case in Section 6.

4. EMPIRICAL EXAMPLE: THE "EXCESS" SENSITIVITY OF CONSUMPTION

Here we present an empirical application that illustrates the ideas discussed in the previous sections: Estimating the elasticity of consumption with respect to anticipated income changes. This application is characterized by a concentration parameter C just above 10. Thus conventional weak instrument testing thresholds are met, but as we will see, issues related to problematic behavior of 2SLS t-tests are still relevant.

Simple versions of the permanent income hypothesis (PIH) imply the consumption elasticity should be zero. A positive value is referred to as "excess sensitivity," which may be evidence of liquidity constraints. But elaborations of the PIH to account for consumption/leisure substitution and/or consumer prudence (reluctance to borrow against uncertain future income) may also generate "excess sensitivity." Regardless, the elasticity of consumption with respect to anticipated income changes is of considerable interest.[10]

To estimate the elasticity we run the regression:

$$\Delta lnC_{it} = \alpha + \beta \Delta lnY_{it} + \boldsymbol{\gamma V}_{it} + \epsilon_{it} \tag{4}$$

where C_{it} is consumption of household i in period t, Y_{it} is household income, and $\mathbf{V}_{it}$ is a vector of control variables. This includes year dummies (to capture business cycle effects). Attanasio and Browning (1995) emphasize the importance of controlling for effects of household demographics on consumption, so we also include age of the household head, the change in age squared, and the change in number of children at home.

To estimate the effect of *anticipated* income changes we need to instrument for ΔlnY_{it} using a variable that is both known to consumers at time $t-1$ and predicts income growth. As Altonji and Siow (1987) pointed out, the instrument must also be uncorrelated with measurement error in income, ruling out using income at $t-1$. Fortunately, income is well approximated by an IMA(1,1) process, so ΔlnY_{it} is MA(2). Following Mork and Smith (1989), this means $lnY_{i,t-2}$ can be used as the instrument for ΔlnY_{it}, as it is known at t-1, predicts income growth, and is uncorrelated with error in measuring ΔlnY_{it} (if measurement error is serially uncorrelated). Following Mariger and Shaw (1993) we test if the MA income process is stable over our sample period, and cannot reject that it is.

We use data from the Panel Study of Income Dynamics (PSID), which follows a sample of over 5,000 U.S. households and their descendants since 1968. We take a subsample of married male household heads aged 23-54 (working age and not near retirement). We use the most comprehensive consumption measure,[11] available from 2005 to 2019. As the PSID became biannual in 1999, we have 8 observations per household. The consumption and income questions refer to the survey year, so in estimating (4) we use changes over two-year intervals. For income, we use total family income, which includes all taxable and transfer income for the head of household, spouse, and any other adults. The use of changes in log consumption and income accentuates measurement error, so as is typical in this literature we introduce a number of data screens to remove outliers.[12] The first of

[10] For instance, if utility is Cobb-Douglas in consumption and leisure then consumption and work hours tend to track closely together in life-cycle models, and consumption drops substantially at retirement.

[11] Total observed consumption is comprised of all food, housing, utilities, transport, education, childcare, healthcare, clothes, vacation, and recreation expenditure.

[12] We restrict the sample to households with income between $\$3,000$ and $\$1,000,000$ in every year. We drop households that report income or consumption changes of less than -70% or more than 300% between any two survey years. We impose a balanced panel by removing households with missing data in any survey year from 2005 to 2019, and drop households if the head has less than 6 years of education.

the 8 observations per household is used to form lagged consumption. This left us with 643 households and 7 observations per household, for a total sample size of $4,501$.

We report the results in Table 3. Estimating (4) by OLS we obtain a coefficient of 0.140 with a standard error of 0.017, indicating a positive covariance between consumption and income changes. But OLS does not estimate the elasticity with respect to anticipated income changes for two key reasons: First, observed income changes include both anticipated and unanticipated components, and the PIH predicts that unanticipated increases in income will increase consumption via an income effect, biasing the coefficient upward. Second, measurement error in income changes is likely to be substantial, biasing the coefficient downward. So the direction of bias is theoretically ambiguous.

Table 3. Elasticity Estimates - PSID

	OLS	2SLS 1^{st} Stage	2SLS 2^{nd} Stage	Reduced Form
Dependent Variable	ΔC_{it}	ΔY_{it}	ΔC_{it}	ΔC_{it}
ΔY_{it}	0.1398 (0.0166) [0.0185]		0.5524 (0.2920) [0.2024]	
ΔlnY_{t-2}		-0.0321 (0.0100) [0.0078]		-0.0177 (0.0085) [0.0062]
F-Stat (Hetero-σ Robust) *p-value*		10.283 0.0014		4.312 0.0379
F-Stat (Cluster Robust)		16.965		8.182
R^2	0.0414	0.0256		0.0224

Note: Heteroskedasticity robust standard errors are in parentheses while standard errors clustered by individual are in square brackets. All regressions control for year effects, age, change in age^2 and change in number of children. $N = 4,501$

We report the first stage 2SLS results in the second column of Table 3. As expected $lnY_{i,t-2}$ is a highly significant predictor of ΔlnY_{it}. Higher income at $t-2$ predicts an income decline from $t-1$ to t, as we expect given the MA(2) structure of ΔlnY_{it}. As we are now using actual data, rather than the *iid* normal data of our sampling experiments, we need to consider robust statistics. The heteroskedasticity robust partial F statistic is 10.28, so it is slightly above the commonly recommended threshold of 10.

The second stage 2SLS result is reported in the third column of Table 3. The estimated elasticity is 0.552, implying OLS is downward biased, and that current consumption is very sensitive to anticipated changes in current income. However, the heteroskedasticity robust standard error is 0.292, so the 2SLS t-test is not significant at the 5% level. In contrast, the last column presents reduced form results. The heteroskedasticity robust partial F statistic is 4.31, so the AR test indicates our elasticity estimate is significant at the 3.8% level. Inverting the AR test,[13] we obtain a 95% confidence interval of (0.03, 1.57)

[13]The inversion requires regressing $y - x\beta_0$ on all exogenous variables, and finding the max and min β_0 values such that the excluded instrument is significant at exactly the 5% level in this regression. We

which excludes 0. Thus, the t-test and AR test disagree. This highlights the question of whether the t-test or AR test is more reliable.

We now investigate the behavior of the AR test and the t-test in this data environment. We conduct the following experiment: Using our PSID sample of N=4,501 observations we can "bootstrap" a new artificial dataset by sampling 4,501 observations with replacement. We do this 10,000 times to form 10,000 artificial datasets. We then repeat the analysis of Table 3 on all 10,000 datasets, and summarize the results in Table 4.

Our method of constructing samples means the point estimates in Table 3 are the true values of the data generating process *in our simulation experiment*,[14] and the concentration parameter C of the DGP is 10.28. Thus, we are above conventional thresholds for an acceptably strong instrument. We begin by noting two features of Table 4:

First, the median OLS, 2SLS and reduced form estimates all agree closely with the point estimates reported in Table 3. The mean OLS and reduced form estimates also agree, while of course the sample mean of the 2SLS estimates does not (as the mean of the 2SLS estimator does not exist in the exactly identified case).

Table 4. Results from Monte Carlo Bootstrap Samples

	OLS		2SLS		F Stat.	Reduced Form	
	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	First Stage	$\hat{\beta}$	S.E.
Median	0.1395	0.0165	0.5502	0.2971	10.3122	-0.0177	0.0085
Mean	0.1395	0.0166	0.6135	2.7765	11.3651	-0.0177	0.0085
Std. Dev.	0.0164	0.0006	1.6630	156.3519	6.6763	0.0085	0.0003

Second, the heteroskedasticity robust standard errors of the OLS and reduced form estimates agree with the empirical standard deviations of those estimates across the 10,000 datasets, and also with the heteroskedasticity robust standard errors reported in Table 3. Thus, the asymptotic standard errors are a good guide to the actual sampling variation of the OLS and reduced from estimates. In contrast, the empirical standard deviation of the 2SLS estimates bears no resemblance to the 2SLS standard error, because the variance of 2SLS does not exist in the exactly identified case.[15]

Now we examine the behavior of the 2SLS standard error. In Figure 5 we plot $se(\hat{\beta}_{2SLS})$ against $\hat{\beta}_{2SLS}$ across the 10,000 samples. A strong *positive* association between 2SLS estimates and standard errors is evident, reversing the pattern in Figure 1. The reversal occurs because in this DGP the correlation ρ between the errors in the structural and reduced form equations is negative (-0.40). As a result, the mean OLS estimate of 0.14 is

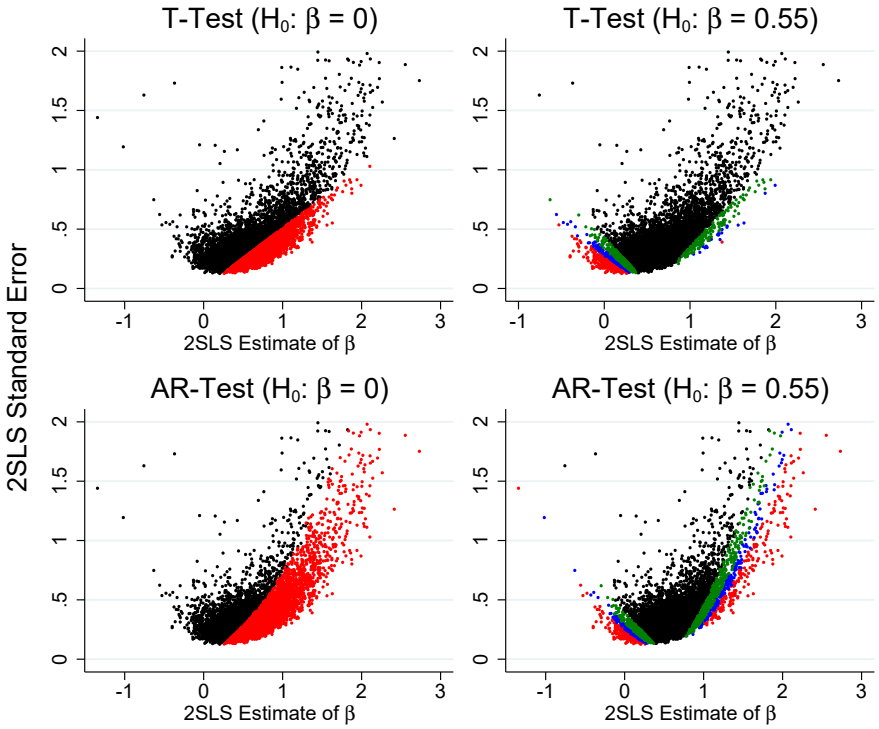
implement this using the Stata command *weakiv* by Finlay and Magnusson (2009), which does a grid search over β_0. It allows for heteroskedastic errors. The Stata command *condivreg* of Mikusheva and Poi (2006) calculates the AR confidence interval analytically, but it assumes homoskedastic errors.

[14]This because the variance-covariance matrix of (y, x, z) in the PSID sample is the population variance-covariance matrix in the simulation experiment.

[15]Table 3 also reports cluster-robust statistics that account for serial correlation. Given the negative serial correlation in residuals induced by the MA structure of consumption changes, this reduces the estimated standard errors. As a result, the cluster-robust 2SLS t-test indicates the elasticity estimate is significant. The cluster-robust standard error is appropriate for applied work in this case, given the panel structure of the data. However, in our simulation experiment we create artificial data by *iid* sampling with replacement from the 4,501 observations. This breaks the panel structure of the data, so the data structure in our sampling experiment is cross-sectional. Hence we focus on the heteroskedasticity robust statistics that ignore serial correlation, as these are what the sampling experiment will mimic.

well below the true elasticity of $\beta = 0.55$. When the OLS bias is *negative* the association between 2SLS estimates and standard errors is *positive*. As we see in Fig. 5, the 2SLS standard errors imply the 2SLS estimates are much more precise when they are in the vicinity of the OLS bias than when they are near the true value of $\beta = 0.55$.

Figure 5. Standard Error of $\hat{\beta}_{2SLS}$ plotted against $\hat{\beta}_{2SLS}$ itself



Note: Runs with standard error > 2 *are not shown. In the left panel, red dots indicate* $H_0 : \beta = 0$ *is rejected at the 5% level, while in the right panel red dots indicate* $H_0 : \beta = 0.55$ *is rejected at the 5% level. Blue and green indicate rejections at the 10% and 20% levels.*

In the top panel of Figure 5 we assess the performance of the t-test. In the top left panel we shade in red cases where $\hat{\beta}_{2SLS}$ is significantly different from zero according to a two-tailed 5% level test. This occurs 39.7% of the time, which is the power level. In the right panel the red dots indicate cases where we reject the true null $\beta = 0.55$. This occurs in 3.58% of cases, so the size of the test is too small. More importantly, almost all rejections occur when $\hat{\beta}_{2SLS}$ is near zero, because the 2SLS standard errors are relatively small when the estimate is shifted in the direction of the OLS bias.

The bottom panel of Figure 5 assesses the performance of the AR test. In the bottom left panel we shade in red cases where $\hat{\beta}_{2SLS}$ is significant at the 5% level, which occurs 54.8% of the time. Thus the AR test exhibits substantially better power than the t-test (54.8% vs. 39.7%). In the right panel we consider AR tests of the true null $\beta = 0.55$. This is simply the (heteroskedasticity robust) partial F-test from a regression of $y - x\beta$ on the instrument and other exogenous variables. The red region again highlights rejections at the 5% level. This occurs in 4.69% of cases, so the size of the test is quite accurate. Moreover, those rejections are almost evenly distributed between cases where $\hat{\beta}_{2SLS}$ is

above vs. below the true value of 0.55. Thus, the AR test largely solves the problem of asymmetry in test results that affects the 2SLS t-test.[16]

These results show that the AR test exhibits both substantially better power and more accurate size than the t-test in this data environment. Moreover, it does not suffer from the problem that estimates shifted in the direction of the OLS bias appear to be more precise. This illustrates that the problems with 2SLS t-tests and advantages of the AR test that we discussed in Sections 2–3 are not limited to the *iid* normal environment, but are also evident in a non-normal environment constructed from actual data.

Based on our experiment, we conclude the AR test should be viewed as more reliable than the t-test in this context. The AR test indicates our elasticity estimate of 0.552 is significant at the 3.8% level, so we gain confidence in that result.

In general, the performance of the t-test deteriorates relative to the AR test as the endogeneity problem becomes more severe. In Keane and Neal (2022a) we present an application to estimating the Frisch labor supply elasticity. In that case $\rho = -0.70$, and the advantages of the AR test are even greater than we see here.

5. CONDITIONAL T-TESTS: THE ACT AND TF TESTS

As an alternative to adopting AR, some authors propose to "fix" the t-test by adjusting its critical values. Conventional critical values rely on the assumption the test statistic is distributed $N(0,1)$ under the null. In fact the 2SLS t-test is highly non-normal, generating size distortions. A "conditional" t-test adjusts the standard t-test critical values to take this non-normality into account, thus obtaining a test with correct size.

5.1. The Asymmetric Conditional t-Test (ACT)

Given standard critical values, 2SLS t-test size depends on C and ρ. Mills et al. (2014) use this relationship to obtain *conditional* critical values that give correct size. This requires simulating the conditional distribution of the t-test, as we show in Appendix B.

Table 5 gives summary statistics of critical values simulated from the DGP in (1) with ρ=0.80. For various levels of C we report the median and standard deviation of the simulated critical values. For example, if C=2.30 ($F_{.05}$=10), which is often considered a standard for an acceptably strong instrument, the median critical values for 2.5% left and right-tailed t-tests are -0.443 and 3.115, respectively. Using these critical values, the one-tailed t-tests have *approximately* the correct 2.5% size. The large deviation from the usual $\pm$1.96 illustrates the extreme power asymmetry of the conventional 2SLS t-test.

As Mills et al. (2014) note, left and right-tail conditional critical values can be combined to form two-tailed conditional t-tests with approximately correct size. For example, we can combine 2.5% left and right-tail critical values to form a two-tailed t-test with approximate size of 5%. We will call this an "asymmetric" conditional t-test (ACT).

In Table 6 we compare power of the ACT and AR tests. We consider true values of β=0.30 or β=-0.30. Clearly, the power of the two tests is almost identical.[17] In the strong instrument case of C=74 both tests exhibit a clear power asymmetry: a 91% rejection

[16] We also shade the 10% and 20% level rejections in blue and green. The AR test rejects at 9.85% and 19.9% rates, so size is accurate, and rejections are evenly distributed above/below the true value. The t-test, in contrast, only rejects at 7.1% and 13.9% rates, with 6.9% and 11.7% in the negative direction.

[17] Andrews et al. (2007) find two-tailed conditional t-tests have very poor power. This is because – unlike the ACT test – the tests considered by Andrews et al. (2007) constrain the critical values to be symmetric around zero, which fails to deal with the power asymmetry problem we have emphasized.

Table 5. Critical Values for Conditional One-Tailed t-tests, $\rho = 0.80$

	1%	2.5%	5%	95%	97.5%	99%
$C = 2.3$	-0.444	-0.443	-0.438	2.564	3.115	3.751
	(0.178)	(0.178)	(0.178)	(0.147)	(0.147)	(0.157)
$C = 10$	-0.925	-0.913	-0.884	2.236	2.762	3.393
	(0.178)	(0.167)	(0.146)	(0.102)	(0.128)	(0.157)
$C = 73.75$	-1.795	-1.585	-1.382	1.885	2.298	2.795
	(0.049)	(0.035)	(0.026)	(0.033)	(0.044)	(0.062)
$C = 336.3$	-2.084	-1.788	-1.524	1.760	2.123	2.554
	(0.032)	(0.024)	(0.019)	(0.025)	(0.032)	(0.046)
$C = 1,000$	-2.187	-1.861	-1.575	1.712	2.055	2.460
	(0.033)	(0.024)	(0.020)	(0.023)	(0.030)	(0.042)

Note: The standard deviations in parentheses are across 10,000 simulations.

rate when β=-0.3 but only a 53% rejection rate when $\beta = 0.3$. As we explain in Appendix A.1, this is because the error variance in the reduced form for y increases with β.

Thus both ACT and AR have good power to detect true effects (β=-0.3) opposite in sign to the OLS bias. Detection of true negative effects when selection into treatment is positive is a top priority if one wishes to adopt a "first do no harm" approach to policy evaluation. The AR and ACT tests are clearly superior to the t-test in this respect.

Table 6. Power of the ACT and AR tests ($\rho = 0.8$) (%)

C	$F_{5\%}$	Conditional t-test (ACT)			AR Test		
		H_0:$\beta = 0$	$\hat{\beta} > 0$	$\hat{\beta} < 0$	H_0:$\beta = 0$	$\hat{\beta} > 0$	$\hat{\beta} < 0$
True $\beta = 0.3$							
2.30	10	7.3	5.6	1.7	6.6	6.5	0.1
5.78	16.38	9.4	8.1	1.3	8.4	8.3	0.2
29.44	50	25.5	25.5	0.1	25.2	25.2	0.0
73.75	104.7	53.7	53.7	0.0	53.4	53.4	0.0
True $\beta = -0.3$							
2.30	10	5.5	0.7	4.8	9.0	3.2	5.9
5.78	16.38	10.7	0.2	10.5	15.0	0.8	14.2
29.44	50	54.8	0.0	54.8	54.7	0.0	54.7
73.75	104.7	91.0	0.0	91.0	91.0	0.0	91.0

Note: The table reports the frequency of rejecting the false null hypothesis H_0:β=0. Columns labeled $\hat{\beta}$>0 and $\hat{\beta}$<0 show how many rejections occur when $\hat{\beta}$ is positive or negative.

Consider now the C=2.30 ($F_{.05}$=10) case, often considered a standard for an acceptably strong instrument. The power of both ACT and AR is very poor. For example, for ACT the probability of rejecting H_0:β=0 is only 7.3% if true β=0.3, and only 5.5% if true β=-0.3. Moreover, many of these rejections occur when $\hat{\beta}_{2SLS}$ has the wrong sign. The obvious conclusion is there is simply not much information in the data, and no choice of testing procedure will change that. This lack of power is concerning given the prevalence of the $\hat{F} > 10$ rule of thumb in practice. We argue that applied researchers should adopt a higher threshold for acceptable instrument strength - see Keane and Neal (2022b).

In the case of C=5.78 ($F_{.05}$=16.38) we see small improvements for both tests. Power attains levels of 8.4% to 15%, and wrong sign rejections become rare. But these power

levels still seem uninspiring. AR has better power to detect a negative β than ACT, 15% vs. 10.7%. Only with C=29.4 ($F_{.05}$=50) does power to detect a true negative exceed 50%.

Overall, we conclude that the ACT test is not a very useful alternative to the AR test in the single instrument case. It is much more difficult to implement and yields very similar results. We revisit this question in Section 6 on the over-identified case.

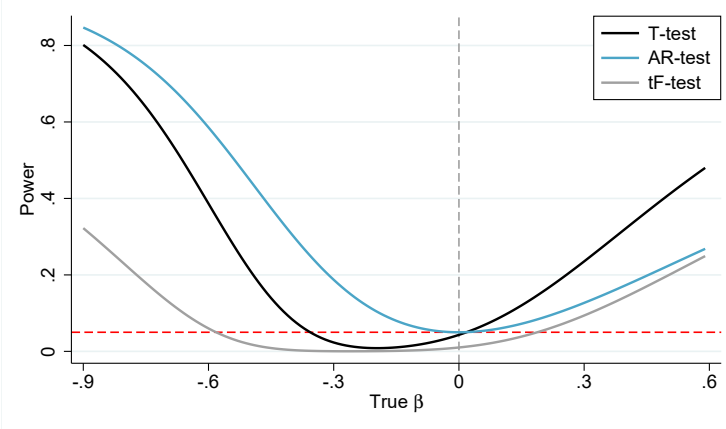
5.2. The tF-test

Lee et al. (2022) propose to eliminate the maximal size distortion of the two-tailed 2SLS t-test by conditioning its critical values on the first-stage $\hat{F}$. They call this the tF-test. It is closely related to the ACT test. The difference is that tF-test critical values are symmetric about zero, and worst-case values are assumed for both ρ and C.

They show a first-stage $\hat{F}$ of 104.7 is required to guarantee size of a 5% two-tailed t-test is no greater than 5% (i.e., worst-case size inflation is zero).[18] Hence, if the first stage $\hat{F} \geq 104.7$ the tF test uses the conventional ± 1.96 critical values. At smaller values of $\hat{F}$ the t-test size is inflated. Hence the critical values must be scaled up to compensate. For example, if $\hat{F} = 10$ one must scale the 5% critical values up to ± 3.43 to reduce the maximum size of the t-test to exactly 5%. The smaller is the first-stage $\hat{F}$, the greater is the required scaling up of the critical values to eliminate size distortion.

When $\hat{F} < 3.84$ both AR and tF 95% confidence intervals are unbounded. The reason is simple: If $\hat{F} < 3.84$ we lack 95% confidence the model is identified. It is logically inconsistent to place a 95% confidence interval on β in this case. Yet a 2SLS t-test based confidence interval does exactly that. The Stock-Yogo approach avoids this problem by requiring one to accept the null if $\hat{F}$ falls below a threshold even higher than 3.84.

Figure 6. Power of the tF, t and AR Tests (C = 10, $\rho = 0.5$)



By construction tF-test critical values are always greater than or equal to conventional t-test critical values. So power of the tF test is unambiguously less than that of the t-test. This can be observed in Figure 6, which compares the power curves of the tF, t and AR

[18]The size of the t-test is increasing in $|\rho|$, with maximal size inflation at $\rho = \pm 1$. Size inflation also depends on C. Lee et al. (2022) show the worst case for C is $[\hat{F}/(\hat{F}^{1/2}+1.96)]^2$. Using a procedure similar to that described in Appendix B, it is possible to simulate the distribution of the t-test conditional on $\hat{F}$, assuming ρ=1 and fixing C at the worst-case level. Using a modified version of Appendix A equation (A2), Lee et al. (2022) show $\hat{F} \geq 104.7$ is required to guarantee the size is no greater than 5%.

tests in the case of $C = 10$ and $\rho = 0.5$. The tF test has low power in general, and very little power to detect true negative effects when the OLS bias is positive. Given the poor power of the tF test relative to AR, we do not advise adopting it when AR is available.

6. THE CASE OF MULTIPLE INSTRUMENTS

Here we consider the over-identified case. As we will see, using multiple instruments increases efficiency but worsens median bias in 2SLS estimates, as well as size distortion and power asymmetry in 2SLS t-tests. This makes the use of alternative estimators and robust test statistics even more important. The LIML estimator is particularly attractive.

With K instruments $\mathbf{z} = (z_1, ..., z_K)$ the definition of the concentration parameter C is unchanged. But population F is C/K and the first-stage sample $\hat{F}$ is $(N/K)\hat{Var}(\mathbf{z\pi})/\hat{\sigma}_e^2$ which has a non-central $F(K, N; C)$ distribution. Table 7 lists, in the $K = 3$ case, several different levels of C, the associated population F, and the 5% critical value of the $F(3, \infty; C)$ distribution to which we compare $\hat{F}$ to test if F exceeds that value.

The case of three instruments is interesting, as K=3 is required for the mean and variance of the 2SLS estimator to exist. We begin with this case, turning later to examine larger instrument sets. We continue to work with the model in (1), and we focus on the simple case where the three instruments are independently distributed $N(0, 1)$, and the π coefficients on the three instruments are equal (so each is equally strong). Table 7 gives examples of levels of C and the maximal size that each achieves.

Table 7. First-Stage $\hat{F}$ Critical Values (K=3)

C	Pop F	π	$\hat{F}_{.05}$ threshold	Max Size
6.90	2.30	0.0480	6.93	28.7%
13.01	4.34	0.0659	10.00	18.6%
40.91	13.64	0.1168	22.30	10%
110.55	36.85	0.1920	50.00	6.8%
360.26	120.09	0.3465	142.50	5.5%

Note: The instrument vector is significant at the 5% level in the first stage if F>2.60. Higher levels of $\hat{F}$ reduce size distortions to levels in column 5. The listed π levels are specific to N=1000.

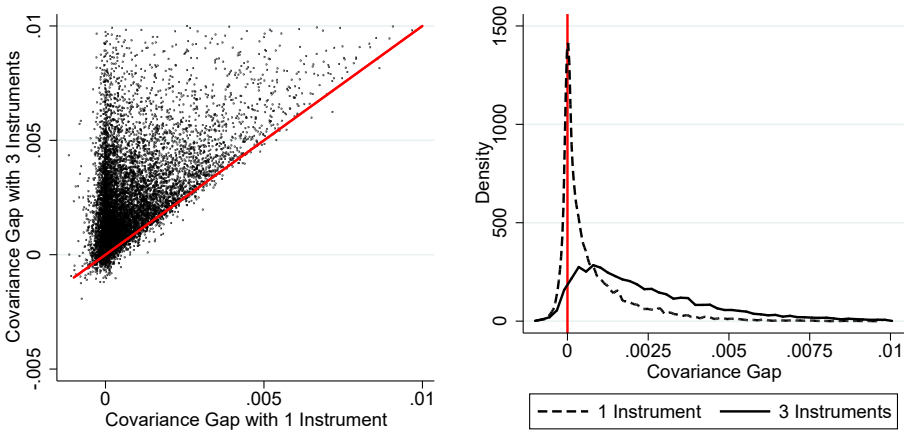
Recall that in the single instrument case π=0.048 and N=1000 gives C=F=2.3. As we noted in Table 1, the associated 5% critical value of the $F(1, \infty; 2.3)$ distribution is 10. Suppose now we have three equally strong instruments. As we see in Table 7 this triples C to 6.9, and leaves population F unchanged at 2.3. The 5% critical value of the $F(3, \infty; 6.9)$ distribution to test that population F is at least 2.3 is now 6.93.

One would think three equally strong independent instruments are better than one, but a problem arises: 2SLS is IV using $\mathbf{z}\hat{\boldsymbol{\pi}}$ as the instrument for x, where $\hat{\boldsymbol{\pi}}$ is obtained from OLS regression of x on $\mathbf{z}$. Many problematic properties of 2SLS arise due to sample covariance of the feasible instrument $\mathbf{z}\hat{\boldsymbol{\pi}}$ with the structural error u. This tends to exceed the covariance of the optimal instrument $\mathbf{z\pi}$ with u, by virtue of how OLS forms $\hat{\boldsymbol{\pi}}$. Unfortunately, the use of multiple instruments worsens the problem: An instrument z_k that happens to have high sample covariance $\hat{cov}(z_k, e)$ with the first-stage error e will tend to get a larger $\hat{\pi}_k$. This drives up the sample covariance $\hat{cov}(\mathbf{z}\hat{\boldsymbol{\pi}}, e)$. And if e and u are correlated (i.e., if we have endogeneity) this also drives up the magnitude of $\hat{cov}(\mathbf{z}\hat{\boldsymbol{\pi}}, u)$.

Figure 9 illustrates this problem. We first define the "covariance gap" as the difference $\hat{cov}(\mathbf{z}\hat{\boldsymbol{\pi}}, u) - \hat{cov}(\mathbf{z\pi}, u)$. Using 10,000 artificial data sets generated from model (1) with

$\rho = 0.5$ and N=1000, we calculate this covariance gap for the cases of 3 vs. 1 instrument. The left panel of Figure 9 shows how the gap almost always increases, often substantially, in the K=3 case. The right panel shows how the density of the covariance gap shifts sharply to the right. Thus, using three instruments increases the sample covariance of $\mathbf{z}\hat{\boldsymbol{\pi}}$ with the structural error, which tends to bias 2SLS estimates towards OLS.

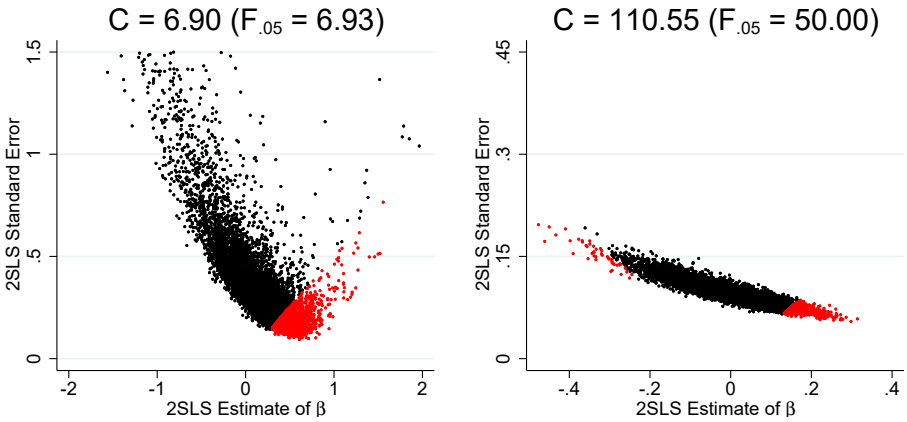
Figure 7. Instrument Endogeneity with One vs. Three Instruments (C=6.9, ρ=0.5)



Note: We define the covariance gap $= \hat{cov}(\mathbf{z}\hat{\boldsymbol{\pi}}, u) - \hat{cov}(\mathbf{z}\boldsymbol{\pi}, u)$ *for the cases of K=1 and K=3. We plot their joint distribution (left), and their marginal densities (right).*

The increased sample covariance $\hat{cov}(\mathbf{z}\hat{\boldsymbol{\pi}}, u)$ strengthens the association between 2SLS estimates and standard errors in the multiple instrument case. This is shown in Figure 8. The left panel is comparable to the left panel of Figure 1, where ρ=0.8 and C=2.3, except now we add two equally strong independent instruments so C=6.9. This causes the Spearman r_s to increase in magnitude from -.576 to -.781. As a consequence, the size of the two-tailed 5% 2SLS t-test increases from 10% to 19.9%. As before, all rejections occur when $\hat{\beta}_{2SLS} > 0$, so only estimates shifted towards the OLS bias are ever significant.

Figure 8. $se(\hat{\beta}_{2SLS})$ plotted against $\hat{\beta}_{2SLS}$ with Three Instruments ($\rho = 0.80$)



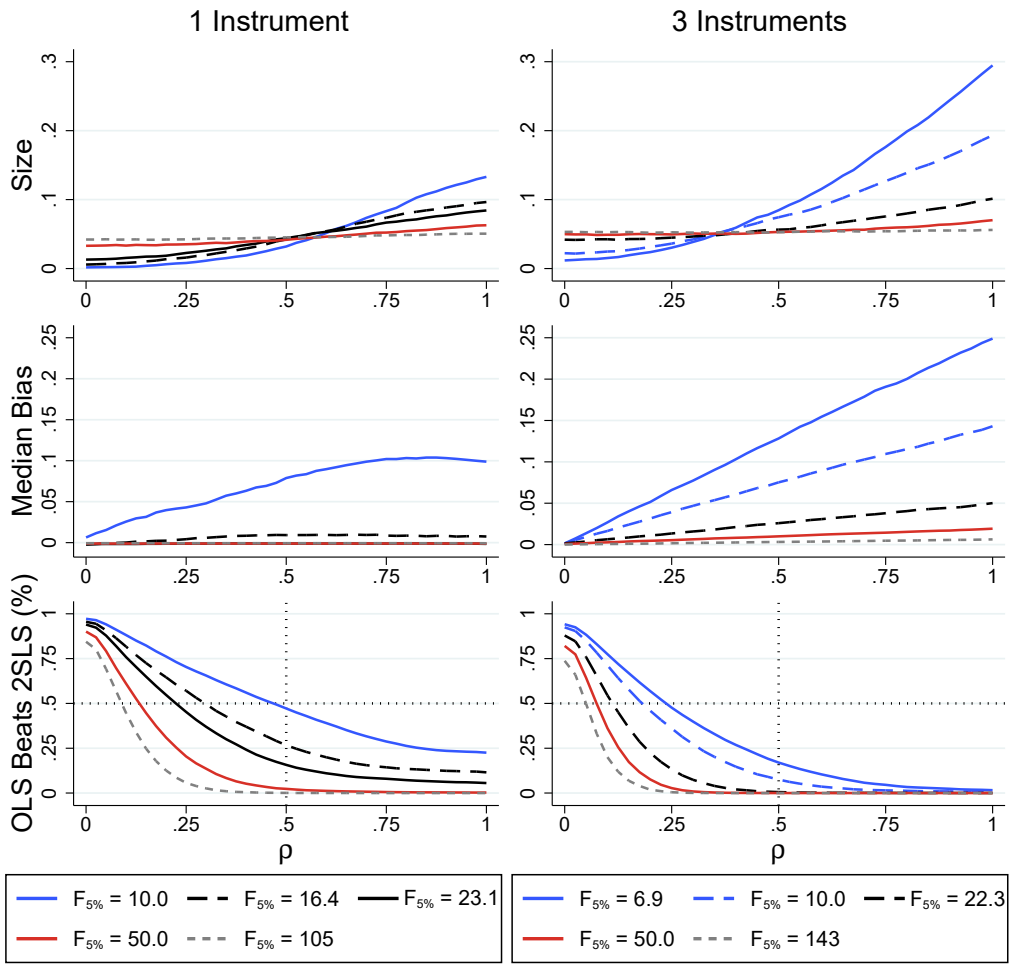
Note: Runs with std. error > 1.5 *not shown. Red dots indicate* $H_0 : \beta = 0$ *rejected at 5% level.*

Increasing instrument strength to the much higher level of C=110.6 ($F_{.05}$=50) does not solve this problem. As we see in the right panel of Figure 8, the Spearman r_s between the 2SLS estimates and their standard errors increases to -.906. The t-test rejection rate is now 6%, so the size distortion is mostly eliminated. But fully 92% of those rejections occur when $\hat{\beta}_{2SLS}$ >0, so the power asymmetry is still severe.

We give a broad overview of the impact of moving from 1 to 3 instruments in Figure 9. The top two panels consider size and median bias. The third panel considers efficiency.

The top panel of Figure 9 plots how size of two-tailed 5% t-tests depends on C and ρ. It is interesting to compare the case of K=1 C=2.3 in left panel with K=3 C=6.9 on the right. These are the two blue lines. This comparison corresponds to adding two new independent instruments of equal strength. With 3 instruments the rejection rate rises much more steeply with ρ, and it peaks at almost 30%, compared to only 13% in the one instrument case. So adding instruments clearly worsens the maximal size distortion.

Figure 9. Properties of 2SLS Estimates and t-Tests with 1 vs. 3 Instruments



Why does t-test size inflation increase with ρ? If instruments are weak and ρ=0 the t-test has little power. So samples where $\hat{\beta}_{2SLS}$ is significant are rare if $\beta = 0$ (i.e., size is less than 5%). But if ρ>0 the power asymmetry problem emerges: A positive sample covariance $\hat{cov}(\mathbf{z}\hat{\boldsymbol{\pi}}, u)$ generates both an estimate shifted towards OLS and a low standard error. As ρ increases this negative association between 2SLS estimates and standard errors gets stronger. Hence, despite $\beta = 0$, as ρ increases we get more samples where $\hat{\beta}_{2SLS}$ is positive and significant, as illustrated in Figures 1-8.

Adding instruments worsens t-test size inflation by increasing the sample covariance between $\mathbf{z}\hat{\boldsymbol{\pi}}$ and u, which, in turn, amplifies the power asymmetry problem at any given level of ρ. Intuitively, the instrument $\mathbf{z}\hat{\boldsymbol{\pi}}$ appears spuriously strong in samples where it is highly correlated with u, and hence with the endogenous part of x. Unfortunately, such samples also tend to generate estimates shifted towards OLS, which generates what we call spurious power. Both higher ρ and more instruments magnify this problem.

With multiple instruments it is far more challenging to satisfy the Staiger-Stock rule of thumb. When K=3, $\hat{F}$>10 gives 95% confidence that C is at least 13. This can be achieved using three independent instruments each with π=0.0659. Compare to the single instrument case: There $\hat{F}$>10 gives 95% confidence that C is at least 2.3, and to achieve this we only need a single instrument with π=0.0480.

In Figure 9, the dotted blue line in the upper right panel shows the K=3 C=13 case, while the solid blue line in the left panel shows the K=1 C=2.3 case. As we see, moving from a single instrument to three independent instruments that are individually 50% stronger actually worsens size inflation considerably.[19]

Thus, if size inflation in two-tailed t-tests is one's primary concern, it is hard to justify using multiple instruments. This is consistent with Angrist and Pischke (2008)'s advice that applied researchers should choose their one best instrument. The same point applies to median bias. The middle panel of Figure 9 shows how median bias varies with C and ρ. Again, it is useful to compare the cases of K=1 C=2.3 and K=3 C=6.9, the two blue lines. This comparison corresponds to adding two new independent instruments of equal strength. Clearly, using more instruments makes median bias unambiguously worse.

We argue, however, it is a mistake to focus exclusively on median bias and size distortions of t-tests in assessing the performance of 2SLS. Efficiency and power are also important. To explore efficiency, the bottom panel of Figure 9 shows how the probability of 2SLS performing worse than OLS varies with C and ρ. We plot the proportion of simulated datasets where $|\hat{\beta}_{2SLS} - \beta| > |\hat{\beta}_{OLS} - \beta|$. As before, it is interesting to compare the cases of K=1 C=2.3 and K=3 C=6.9, the two blue lines.

We see the addition of two equally strong independent instruments tremendously increases the probability that $\hat{\beta}_{2SLS}$ is closer to the truth than $\hat{\beta}_{OLS}$. There is obviously a large efficiency gain from using the additional information. This gives a very different perspective on the potential efficacy of using multiple instruments. In the next section we ask if alternative estimators or robust tests can successfully exploit the information in multiple instruments without creating the problems that afflict 2SLS and the t-test.

Finally, a notable aspect of the bottom panel of Figure 9 is the high frequency with which the OLS estimates are closer to the truth than 2SLS, particularly in the single

[19]Stock and Yogo (2005) show that C=40.9 ($F_{.05}$=22.3) achieves a maximum size of 10% in the three instrument case. Figure 9 shows this is accurate. This corresponds to a π of 0.1168 on each of the three instruments in the first stage (see Table 7). But as we saw in Table 1, the same objective could be achieved just by using a single instrument with π=0.076.

instrument case. Consider the C=2.30 ($F_{.05}$=10) case, often considered a standard for an acceptably strong instrument. Only as ρ approaches 0.50 does the probability that $\hat{\beta}_{2SLS}$ is closer to the truth than $\hat{\beta}_{OLS}$ pass 50%. And even as ρ approaches 1.0 the probability that 2SLS will outperform OLS barely passes 75%.

Consider now the case of C=29.4 ($F_{.05}$=50). At this higher level of instrument strength the chance that $\hat{\beta}_{2SLS}$ is closer to the truth than $\hat{\beta}_{OLS}$ nears 100% as ρ approaches 0.50. Furthermore, as we see in Figure 9, t-test size distortions are minor, even if K=3, regardless of the level of ρ. The same is true for median bias. Based on such considerations, we think a compelling case can be made for changing the rule of thumb for acceptable instrument strength to $\hat{F} > 50$. We present further arguments in Keane and Neal (2022b).

We emphasize however, that even at C=29.4 ($F_{.05}$=50) the t-test exhibits severe power asymmetry, and has little power to detect large negative effects that are opposite in sign to the OLS bias. So it is still important to use robust tests like AR in lieu of the t-test.

6.1. The LIML Estimator and the LR Test

In the over-identified case the limited information maximum likelihood (LIML) estimator of Anderson and Rubin (1949) is an important alternative to 2SLS, and the associated likelihood ratio (LR) test is an important alternative to both the AR and t-tests. The typical exposition of LIML in textbooks is dauntingly complex, which perhaps explains why applied researchers rarely use it. In fact the idea is very simple:

Say we wish to estimate the parameter β in a simple regression of y on x. Let $\hat{\beta}$ denote a candidate estimate, and consider the residuals $\hat{u} = y - \hat{\beta}x$. Imagine that after choosing $\hat{\beta}$ we run an auxiliary regression of the residuals $\hat{u}$ on x. OLS chooses $\hat{\beta}$ so the R^2 of this regression is exactly zero, which is equivalent to setting $cov(x, \hat{u})$=0. Similarly, 2SLS given a single instrument z can be understood as choosing $\hat{\beta}$ so the R^2 of the regression of the residuals $\hat{u}$ on z is exactly zero, thus setting $cov(z, \hat{u})$=0.

Now say we have a vector of $K > 1$ instruments $\mathbf{z} = (z_1, ..., z_K)$. We would like to adopt a similar strategy, but with multiple instruments it is generally impossible to find a $\hat{\beta}$ that sets the R^2 from regressing the residuals $\hat{u} = y - \hat{\beta}x$ on $\mathbf{z}$ to exactly zero. To do so we would need to set $cov(z_k, \hat{u}) = 0$ for each of the K instruments, but that is generally impossible as we only have a single $\hat{\beta}$ to work with. We are trying to solve K equations with one unknown.[20] What is the solution?

The LIML estimator solves the problem in a simple and natural way. It chooses $\hat{\beta}$ to set the R^2 from regressing the residuals $\hat{u} = y - \hat{\beta}x$ on $\mathbf{z}$ as close to zero as possible. This is achieved by solving a standard problem in linear algebra – the generalized eigenvalue problem. One doesn't need to know the details to understand how LIML works, just as one doesn't need to remember Gaussian elimination to understand OLS.

The 2SLS estimator solves this problem in a different way: As it is not possible to set the R^2 in the residual regression to zero, 2SLS gets around the problem by condensing the K instruments into the single instrument $\hat{x} = \mathbf{z}\hat{\boldsymbol{\pi}}$, where $\hat{\pi}$ is obtained from the first-stage regression of x on $\mathbf{z}$. It then chooses the $\hat{\beta}$ that sets the R^2 from regressing the residuals $\hat{u} = y - \hat{\beta}x$ on the single instrument $\hat{x} = \mathbf{z}\hat{\boldsymbol{\pi}}$ exactly equal to zero.

Some comments are in order: The advantage of the 2SLS approach is computational. The linear algebra for solving a linear system of equations is slightly simpler than the

[20] If all instruments are valid – i.e., the population $cov(z_k, u) = 0$ for each instrument – it should be possible to choose $\hat{\beta}$ so the R^2 is small. But we can never make it exactly zero due to sampling variation.

eigenvalue problem. But that is hardly relevant on modern computers. The downside of 2SLS is that the first-stage tends to put greater weight (larger $|\hat{\pi}_k|$) on individual instruments that have greater sample correlations with the structural error. This induces bias and size distortion. As we will see below, LIML is less subject to these problems.

Second, when K=1, the same $\hat{\beta}$ sets $R^2 = 0$ in regressions of the residuals $\hat{u}$ on either $z\hat{\pi}$ or z (both scalars). Thus 2SLS and LIML are identical in the exactly identified case.

Third, theorists often advise the use of LIML over 2SLS in contexts with many weak instruments. However, when covariances between the instruments $\mathbf{z}$ and the endogenous variable x are low, the R^2 from regressing the $\hat{u} = y - \hat{\beta}x$ on $\mathbf{z}$ may be approximately minimized over a wide range of $\hat{\beta}$. This may cause LIML to "blow up" when instruments are very weak. We suspect this behavior has often caused applied researchers to abandon LIML, when in fact it should be viewed as a warning sign of weak identification.

Fourth, textbook expositions of LIML often define the estimator as the solution to an eigenvalue problem, without motivating (as we did above) why this problem is interesting. We suspect this is a key reason LIML has not caught on with applied researchers.

Fifth, the R^2 from regressing the residuals $\hat{u} = y - \hat{\beta}x$ on on $\mathbf{z}$ has the form:

$$R^2_{\hat{u},\mathbf{z}} = \frac{\hat{U}'Z(Z'Z)^{-1}Z'\hat{U}}{\hat{U}'\hat{U}} \tag{5}$$

where $\hat{U}$ and Z are $N \times 1$ and $N \times K$ vectors that stack the N individual observations. This has an R^2 form as the denomination is the total sum of squares of $\hat{u}$, while the numerator is the variance in $\hat{u}$ explained by projection on $\mathbf{z}$. Solving for the $\hat{\beta}$ that minimizes (5) requires solving a generalized eigenvalue problem. That is because both the numerator and denominator depend on $\hat{\beta}$, via the residuals $\hat{u} = y - \hat{\beta}x$.

Sixth, if we multiply the $R^2_{\hat{u},\mathbf{z}}$ in (5) by N we obtain the "Sargan statistic." It is distributed $\chi^2(K-1)$ in large samples if the instruments are valid (i.e., $cov(z_k, u) = 0 \forall k$). Of course the Sargan statistic increases mechanically as we add valid instruments, for the usual reason that R^2 always increases as we add (irrelevant) variables to a regression. This is why the degrees of freedom of the statistic increase with K. A "surprisingly" large value of the Sargan statistic calls into question the validity of the instruments. Obviously the LIML estimator minimizes the Sargan statistic.

Seventh, the 2SLS estimator chooses $\hat{\beta}$ to minimize the numerator of the Sargan statistic, while ignoring the denominator. This simplifies the eigenvalue problem to the slightly simpler problem of solving a linear system of equations. The two-step GMM estimator (GMM-2S) treats the denominator of the Sargan statistic as given, so under homoskedasticity it solves the exact same problem as 2SLS and is equivalent. The Continuously Updated GMM estimator (GMM-CU) of Hansen et al. (1996) treats the denominator as a function of $\hat{\beta}$, so under homoskedasticity it is equivalent to LIML.

Having implemented LIML, one can evaluate the significance of the estimate using the likelihood ratio (LR) statistic. One can also form a t-test by plugging the LIML residuals into the conventional standard error and t-stat formulas. But the LR test has important advantages, as we will see below. The LR test is based on the reduced form:

$$\begin{aligned} y &= \boldsymbol{z}(\beta\boldsymbol{\pi}) + v \\ x &= \boldsymbol{z}\boldsymbol{\pi} + e \end{aligned} \tag{6}$$

Here $v = \beta e + u$ and, as we suspect $cov(e, u) \neq 0$, the errors in the two reduced form equations are potentially correlated. Thus we can estimate the reduced form as a system

of two seemingly unrelated regressions (SUR) - see Zellner (1962). The constraint that the coefficients on $\mathbf{z}$ in the y and x equations are proportional (in ratio β) is equivalent to the assumption that the instruments are excluded from the structural equation for y.

An LR test for H_0:β=0 is obtained by running two versions of the reduced form system, one imposing β=0 and one imposing $\beta = \hat{\beta}_{LIML}$. In the first case, $\mathbf{z}$ drops out of the y equation. In the second case $\mathbf{z}$ enters the y equation, but the coefficients are constrained to be $\hat{\beta}_{LIML}$ times the coefficients on $\mathbf{z}$ in the x equation. Mechanically, one can implement this by creating new variables equal to $\mathbf{z}\hat{\beta}_{LIML}$ and entering them in the y equation, and subsequently estimating the SUR system under the constraint that $\hat{\boldsymbol{\pi}}$ is equal in both equations. Under normality, one can form the log likelihood for each model. The LR test for H_0:$\beta = 0$ equals two times the deterioration in the log likelihood when the constraint $\beta = 0$ is imposed, and it is distributed as a $\chi^2(1)$ under the null hypothesis.

The LR test takes a very simple form. As we show in Appendix C, it only depends on the variance-covariance matrix of the residuals in the reduced form. Intuitively, if the residual variance increases substantially when the $\beta = 0$ constraint is imposed this indicates that $\hat{\beta}_{LIML}$ is significant. Estimating the SUR system with the constraint $\beta = \hat{\beta}_{LIML}$ also delivers as a by-product the LIML estimate of $\boldsymbol{\pi}$, although this is often not of primary interest. Regardless, it is always important to report the first-stage $\hat{F}$ from the reduced form regression of x on $\mathbf{z}$, whether one uses LIML or 2SLS.[21]

It is important to understand the difference between the LR and AR tests. Recall that AR is the F-test from regressing y on $\mathbf{z}$. In large samples we may also define AR as the NR^2 from regressing y on $\mathbf{z}$, which is distributed $\chi^2(K)$ under the null $\beta = 0$.[22] The AR test has degrees of freedom K to adjust for the mechanical increase in R^2 that occurs as we add instruments (because R^2 always increases when we add regressors).

In contrast, LR tests the single constraint $\beta = 0$ directly, so it is $\chi^2(1)$ regardless of K. A key fact is that the LR statistic equals AR minus the Sargan statistic. So LR takes the NR^2 from regressing y on $\mathbf{z}$, and subtracts off the "mechanical" part of the NR^2 that arises from regression of $\hat{u}$ on $\mathbf{z}$. Thanks to this adjustment, the degrees of freedom of the LR statistic is one regardless of K. This is more efficient than increasing the degrees of freedom as K increases, so the LR test has better power than AR.

Unlike AR, the LR test is not pivotal when K >1. It uses estimated $\hat{u}$ as input, and these depend on $\hat{\beta}_{LIML}$, whose distribution depends on instrument strength. So unlike AR, the LR test is not guaranteed to have correct size when instruments are weak. The choice between AR and LR involves a trade off between power and size distortion.

Thus Moreira (2003) has developed a conditional likelihood ratio test (CLR) that adjusts the critical value of the LR test based on the first stage $\hat{F}$, so the resulting CLR test has approximately the correct size under the null hypothesis. We explain the CLR test in detail in Appendix C. In the single endogenous variable exactly-identified case, the AR test, LR test, CLR test and Langrange multiplier (LM) test (Kleibergen 2002) are all equivalent. In more general settings these tests differ.

[21] Of course, the OLS estimate of $\boldsymbol{\pi}$ obtained by regressing x on $\mathbf{z}$ (i.e., the first stage of 2SLS) is consistent regardless of the true value of β. The LIML estimate of $\boldsymbol{\pi}$ is also consistent (provided the instruments are valid). In contrast, the SUR system that imposes the constraint $\beta = 0$ delivers a consistent (and efficient) estimate of $\boldsymbol{\pi}$ if and only if the constraint $\beta = 0$ is true. If not, the misspecification of the y equation will impart asymptotic bias to the $\boldsymbol{\pi}$ estimate. Thus a comparison of the constrained SUR estimate of $\boldsymbol{\pi}$ with the OLS estimate of $\boldsymbol{\pi}$ provides a Hausman-type test of H_0:$\beta = 0$. Interestingly, Van de Sijpe and Windmeijer (2022) show this is equivalent to the AR test.

[22] One translates from the χ^2 to the F version of AR simply by dividing by K.

6.2. Performance of Alternative Estimators and Tests

The over-identified case offers a wider choice of estimators (2SLS, LIML, GMM) and tests (t, AR, LR, CLR) than the single instrument case. Here we consider the performance of the main alternatives. We start by considering the three instrument case. This is of particular interest as $K \geq 3$ is required for the mean and variance of 2SLS to exist.

Table 8 compares the performance of the 2SLS estimator combined with the t-test vs. LIML combined with AR, LR or CLR. We consider first a DGP with a high level of endogeneity, ρ=0.80, to test how the procedures perform in a difficult environment. Later we look at smaller ρ. We set true β to 0, -0.3 or +0.3. Recall that $\beta = \pm 0.3$ correspond to fairly large effects, as they imply a one std. dev. change in x induces an 0.25 std. dev. change in y. We consider the 5 levels of instrument strength listed in Table 7.

As the mean of 2SLS now exists, it is possible to analyze how bias varies with C. With K=3, Stock and Yogo (2005) show first-stage $\hat{F} \geq 9.08$ gives 95% confidence C is large enough so worst-case bias of 2SLS is less than 10% of the OLS bias. Similarly, $\hat{F} \geq 13.91$ gives 95% confidence the worst-case bias is less than 5%. Thus, the second level of instrument strength in Table 8, $F_{5\%} = 10$, gives high confidence bias is below 10%, and the third level, $F_{5\%} = 22.3$, is far more than adequate to reduce it below 5%.

Turning to Table 8, the poor performance of 2SLS and the t-test is striking:

First, 2SLS suffers from substantial median bias towards OLS. In the $F_{5\%}$ cases of 6.93 and 10, median bias is .200 and .114. These are quantitatively large values if interpreted as effect sizes. They are 25% and 14% of the substantial OLS bias of $\rho/Var(x) \approx 0.80$.[23]

Second, the 2SLS t-test suffers from large size distortion: In the cases of C=6.9 or C=13 a two-tailed 5% t-test rejects the true null β=0 at 19.8% and 13.4% rates, respectively. Furthermore, <u>all</u> rejections occur when $\hat{\beta}_{2SLS} > 0$, so size distortion in one-tailed t-tests is twice as great. The t-test is biased towards finding positive effects.

Third, and even more striking, when true β=-0.30 the t-test has essentially no power to detect a substantial true negative effect. Remarkably, if C=6.9, the t-test rejects H_0:β=0 only 2.5% of the time, and <u>all</u> the rejections happen when $\hat{\beta}_{2SLS} > 0$. That is, we conclude β is positive when it is actually negative![24]

The superior performance of LIML combined with the AR test is evident in Table 8. LIML essentially eliminates median bias. The size of a 5% AR test is correct. There is a slight power asymmetry when instruments are weak (C=6.9) but it vanishes quickly as instrument strength increases. Importantly, the AR test has much better power than the 2SLS t-test to detect true negative effects. For example, when C=13 and true $\beta = -0.3$ it rejects the false null β=0 at a 18.9% rate compared to only 0.7% for the t-test.

The LR test also performs far better than the t-test. The size distortion that affects the LR test is very modest. If C=6.9 it rejects the true null $\beta = 0$ at a 6.5% rate. The weak IV literature does not tend to regard size distortion as small as 1.5% as a serious problem. For instance, the widely used Stock-Yogo tests assess whether t-test size distortions are less than 5% or 10%. We also see that the LR test has superior power to the AR test.

[23] Interestingly, the Stock-Yogo analysis indicates C=13 ($F_{5\%} = 10$) guarantees *mean* bias of less than 10% of OLS. Here the mean bias is 8.4%, but the *median* bias (14%) is much greater.

[24] Conversely, the t-test judges positive estimates to be significant far too often. Consider the bottom panel, where β=0.30. In the C=6.9 case the t-test rejects H_0:β=0 at a very high 46% rate, despite the weakness of the instrument. As we explained in Section 3, this apparently high level of power arises because 2SLS standard errors are spuriously precise when $\hat{\beta}_{2SLS} > 0$. The substantial median bias of 2SLS towards OLS in the multiple instrument case magnifies this problem.

Table 8. 2SLS vs. LIML: Size and Power with K=3 ($\rho = 0.8$) (%)

C		6.90	13.01	40.91	110.55	360.26
$F_{5\%Crit}$		6.93	10.00	22.30	50.00	142.50
		Reject H_0:$\beta = 0$ when True $\beta = 0$				
2SLS	t-Test	.198	.134	.082	.060	.054
	False Positives	.198	.134	.082	.055	.040
LIML	AR Test	.051	.051	.051	.051	.051
	False Positives	.032	.026	.026	.026	.025
	LR Test	.065	.056	.050	.049	.050
	False Positives	.038	.027	.024	.024	.024
	CLR Test	.050	.049	.049	.048	.049
	False Positives	.029	.024	.023	.023	.024
		Reject H_0:$\beta = 0$ when True $\beta = -0.3$				
2SLS	t-Test	.025	.007	.354	.949	1.000
	False Positives	.025	.006	.000	.000	.000
LIML	AR Test	.120	.189	.521	.936	1.000
	False Positives	.014	.003	.000	.000	.000
	LR Test	.191	.289	.689	.981	1.000
	False Positives	.011	.001	.000	.000	.000
	CLR Test	.162	.268	.683	.980	1.000
	False Positives	.009	.001	.000	.000	.000
		Reject H_0:$\beta = 0$ when True $\beta = 0.3$				
2SLS	t-Test	.460	.453	.583	.838	.997
LIML	AR Test	.075	.097	.221	.540	.977
	LR Test	.115	.150	.333	.712	.993
	CLR Test	.097	.138	.327	.708	.993
				Median Bias		
2SLS		.200	.114	.037	.014	.004
LIML		.007	-.002	-.002	-.001	.000

Note: The table reports the frequency of rejecting the null hypothesis H_0: $\beta = 0$.

For example, when C=13 and true $\beta = -0.3$ the LR test rejects the false null $\beta = 0$ at a 28.9% rate compared to 18.9% for the t-test.

The modest size distortion and superior power of the LR test suggest it may be possible to correct its size distortion while still maintaining superior power to the correctly-sized AR test. The CLR test results in Table 8 show this is true. CLR has approximately correct size by construction. And we see it does have superior power to AR. For example, when C=13 and true $\beta = -0.3$ it rejects the false null $\beta = 0$ at a 26.8% rate compared to only 18.9% for AR (and 0.7% for the 2SLS t-test). Given such results, it is difficult to justify using 2SLS or the t-test when LIML and the CLR test are available.

The above discussion highlights a general point: The AR test, which is pivotal and *always* has correct size, should be viewed as a baseline against which other tests are judged (see Dufour (1997)). The fact that a new test has correct size does not make it valuable unless it has superior power to AR. The CLR correction to the LR test is valuable, as CLR has correct size but still maintains superior power to AR. In contrast,

for example, the tF test correction to the t-test – see Section 5.2 – has correct size but greatly inferior power to the AR test. Thus it does not represent an improvement over AR. Unfortunately, as we noted in the introduction, the heavy focus of the weak IV literature on bias and size has often caused it to ignore power considerations.

The Online Appendix reports results with lower levels of endogeneity. When ρ=0.50 the results are similar. Median bias in 2SLS is still substantial. The 2SLS t-test suffers from less size distortion, but the power asymmetry problem remains serious. The t-test has almost no power to detect true negative effects in the C=6.9 and 13 cases, and clearly inferior power to CLR in the C=40.9 case. The combination of median bias and power asymmetry biases the t-test towards finding positive effects (in the direction of OLS). Remarkably, when ρ=0.20 the power asymmetry in the t-test remains substantial. The size of the 5% t-test now falls well below 5% unless instruments are very strong (C=111). This highlights how the t-test derives much of its power from finite-sample correlation between the instrument and the structural error, which generates spurious correlation between x and z. This source of power is limited when endogeneity is weak.

We now return to the ρ=0.80 case, and compare properties of 2SLS and LIML standard errors. The upper left panel of Figure 10 plots the 2SLS standard error against the 2SLS estimate for the case of C=13 ($F_{5\%}$=10), with true β=0. We see the familiar comet shape that illustrates the strong negative correlation between 2SLS estimates and their standard errors. In the upper left panel the red shaded area shows the 13.4% of cases where a 5% level two-tailed t-test rejects H_0:β=0. Due to the power asymmetry that afflicts the t-test, all of these rejections occur when $\beta > 0$ (i.e., in the direction of the OLS bias). Thus a 2.5% level one tailed t-test of H_0:$\beta \leq$0 rejects 13.4% of the time.

The upper right panel of Figure 10 plots the LIML standard errors against the LIML estimates. A crucial point is the LIML standard error has a strong negative association with the LIML estimate – the same problem that afflicts 2SLS. If one were to form t-tests based on LIML standard errors, there would be a size distortion (in this case the null H_0:β=0 is rejected at a 7.5% rate), and all rejections occur when $\beta > 0$. Thus, the use of LIML alleviates but does not solve the two key problems that afflict the t-test.

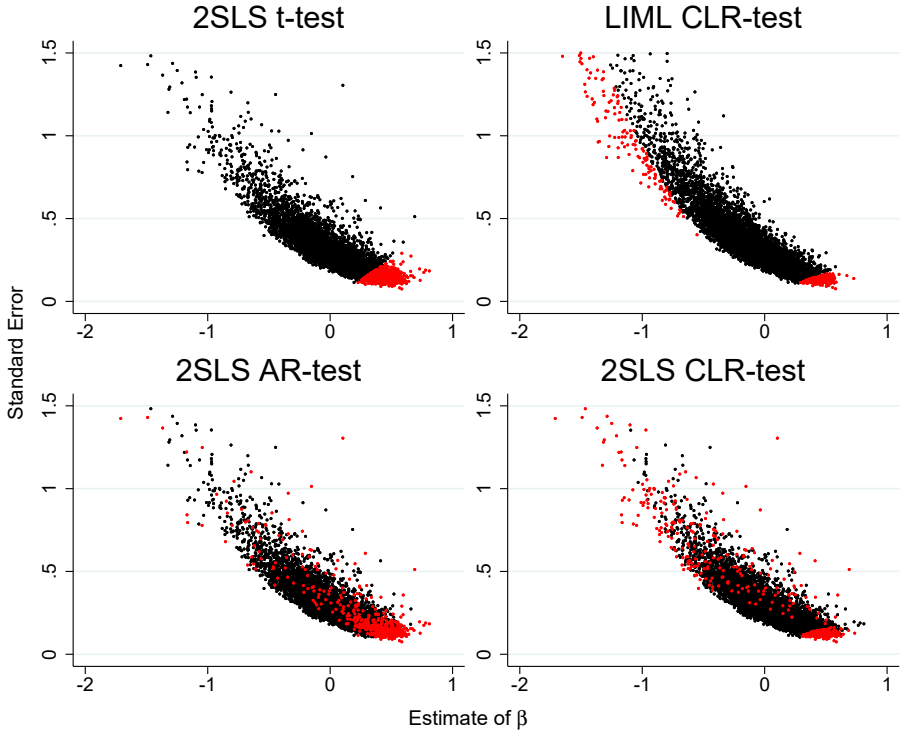
The upper right panel of Figure 10 shows how the use of LIML plus CLR resolves both problems. The red shaded region now shows the approximately 5% of cases where the 5% level CLR test rejects H_0:β=0, so size is correct. And we see these cases are evenly split between positive and negative estimates, so the power asymmetry problem is resolved. It is worth noting at this point that the CLR test may be inverted to form confidence intervals, so the LIML standard error is not needed for that purpose.

Given heteroskedastic data, the GMM-2S and GMM-CU estimators are important options. In Keane and Neal (2022a) we compare all estimators and tests discussed here in an empirical application to estimating the Frisch elasticity. There, we estimate ρ=-0.70, so the power asymmetry problem is substantial. We show that GMM-2S has problems similar to 2SLS, while GMM-CU offers improvements similar to LIML. The advantages of the CLR test over the t-test are substantial. Thus, for heteroskedastic data, we strongly advise using the CLR test in conjunction with either LIML or GMM-CU.

6.2.1. Combining 2SLS with AR, ACT, CLR

The theory literature often advises applied researchers who use 2SLS to report robust tests like AR or CLR if weak instruments are a problem (see, e.g., Stock et al. (2002), Andrews et al. (2019)). This would certainly be an improvement over reporting t-tests.

Figure 10. SE($\hat{\beta}$) plotted against $\hat{\beta}$ itself, $C = 13$ ($F_{5\%} = 10$), ρ=0.80



Note: Runs with std. error > 1.5 *not shown. Red dots indicate* $H_0 : \beta = 0$ *rejected at 5% level.*

But a concern in the over-identified case is that AR and CLR are specifically designed for use with LIML. The bottom panel of Figure 10 illustrates a problem that arises from "mix and matching" estimators and test statistics in this way. It can happen that the 2SLS estimate is near zero, but the AR or CLR test indicates the estimate is significant. As we see in the upper right, the use of LIML with CLR avoids this problem.[25]

Putting this issue aside, we nevertheless consider the performance of 2SLS in conjunction with three alternatives to the t-test – the AR, ACT and CLR tests. The results are reported in Table 9. All three tests are robust to weak instruments - meaning they have correct size regardless of instrument strength – correcting the size distortion of the t-test. So the choice among them can be based on power considerations. On this basis the CLR test is clearly preferred. It has better power than AR or ACT to detect both positive and negative true values of β, regardless of the level of instrument strength.

Of course, a CLR test gives the same result whether the researcher reports the 2SLS or LIML estimate of β. But CLR is based on the LIML estimate, which gives a better match between the hypothesis test result and the magnitude/sign of the parameter estimate. We already saw this in Figure 10. We can also see it by comparing Tables 8 and 9. For example, when true $\beta = -0.3$ CLR rejects $\beta = 0$ 16.2% of the time, but in 2.2% of those cases the 2SLS estimate is positive (the wrong sign) while this problem only arises in 0.9% of cases for LIML. However, this problem vanishes quickly as instrument strength increases and 2SLS and LIML converge.

[25]Of course this issue does not arise in the exactly-identified case as LIML and 2SLS are identical.

Table 9. Alternatives to the 2SLS t-test: Size and Power with K=3 ($\rho = 0.8$) (%)

C		6.90	13.01	40.91	110.55	360.26
$F_{5\%Crit}$		6.93	10.00	22.30	50.00	142.50
		Reject H_0:$\beta = 0$ when True $\beta = 0$				
2SLS	AR Test	.051	.051	.051	.051	.051
	False Positives	.048	.042	.036	.032	.029
	ACT Test	.051	.051	.051	.051	.050
	False Positives	.026	.025	.025	.024	.024
	CLR Test	.050	.049	.049	.048	.049
	False Positives	.036	.028	.023	.023	.024
		Reject H_0:$\beta = 0$ when True $\beta = -0.3$				
2SLS	AR Test	.120	.189	.521	.936	1.000
	False Positives	.046	.023	.002	.000	.000
	ACT Test	.116	.229	.659	.978	1.000
	False Positives	.003	.001	.000	.000	.000
	CLR Test	.162	.268	.683	.980	1.000
	False Positives	.022	.005	.000	.000	.000
		Reject H_0:$\beta = 0$ when True $\beta = 0.3$				
2SLS	AR Test	.075	.097	.221	.540	.977
	ACT Test	.075	.109	.311	.695	.992
	CLR Test	.097	.138	.327	.708	.993

Note: The table reports the frequency of rejecting the null hypothesis H_0: $\beta = 0$.

In the single instrument case the AR test is optimal: It has the best power of any correctly-sized test. As Tables 8-9 illustrate, this optimality property vanishes in the over-identified case, as CLR has better power. The 2SLS + AR combination also suffers from a clear power asymmetry problem: For example, as we see in Table 9, in the weak instrument case of C=6.9 nearly all rejections of the true null β=0 occur when $\hat{\beta}_{2SLS} > 0$. And if true β=-0.3 AR rejects the false null β=0 only 12% of the time, and 39% of those rejections occur when $\hat{\beta}_{2SLS} > 0$, so we often conclude β is positive when it is actually negative. Thus, the AR test is more likely to judge 2SLS estimates significant when they are shifted in the direction of the OLS bias (for reasons we explained in Section 3). This problem is much less severe than it is for the t-test, but it is still of concern.

The AR, LR and CLR tests all have much better power to detect true effects that are *opposite* in sign to the OLS bias, even when instruments are quite strong. For example, in the case of C=41 ($F_{.05}$=22.3) the AR test rejects H_0:β=0 at a 51.1% rate when true β is -0.30, but only a 22.1% rate when true β is 0.30. Similarly, the figures for CLR are 68.3% vs. 32.7%. This occurs because larger β increases the noise in the reduced form relationship between y and z, as we explain in Appendix A.1. The AR test properly interprets this as implying less certainty about significance of $\hat{\beta}$. Recall that LR and CLR are based on AR minus the Sargan statistic, so they follow the same pattern.

Finally, consider the ACT test. If β=0 it has (approximately) the correct 5% size by construction. In contrast to the AR and CLR tests, it also has symmetric rejections on both sides of 0 even when instruments are weak. So if test size were one's only concern the ACT test would be recommended. However, as Table 8 reveals, when $\beta = \pm 0.3$ the ACT test has inferior power to the CLR test, particularly when instruments are weak.

6.3. The Case of Many Instruments

Since the work of Bound et al. (1995), the IV literature has devoted a great deal of attention to the case of many instruments. As we illustrated at the start of this section, going from one to three instruments drives up the sample covariance between the 2SLS instrument $\mathbf{z}\hat{\boldsymbol{\pi}}$ and the structural error u. Using larger numbers of instruments worsens this problem, increasing the bias of 2SLS towards OLS, increasing t-test size distortion, and worsening the power asymmetry of the t-test.

We illustrate the problem in Table 10. We run experiments based on the DGP in (1) with either 10 or 20 independent and equally strong instruments. The table compares the performance of 2SLS combined with t vs. LIML combined with CLR. We consider a high level of endogeneity (ρ=0.80) to test how these procedures perform in a challenging environment. We consider three levels of instrument strength:

In the first case we set π=0.048 for each instrument. Recall from Table 1 that a single instrument of this strength gives C=2.3, and a first-stage $\hat{F}$ of 10 is required to have 95% confidence the instrument is this strong. Using 10 independent instruments of the same strength increases C to 23, and $\hat{F}$ of 5.19 is required to have 95% confidence that C is this large. In the 20 instrument case we have C=46 and $F_{5\%}$=4.61, respectively.[26]

Table 10. Results with Many Instruments ($\rho = 0.8$) (%)

		10 Instruments			20 Instruments		
C		23.00	61.59	314.90	46.00	138.33	1114.00
$F_{5\%Crit}$		5.19	10.00	38.54	4.61	10.00	62.30
		Reject H_0:$\beta = 0$ when True $\beta = 0$					
2SLS	t-Test	.384	.199	.082	.604	.289	.082
	False Positives	.384	.199	.078	.604	.289	.076
LIML	CLR Test	.049	.047	.048	.054	.053	.054
	False Positives	.026	.025	.024	.028	.028	.029
		Reject H_0:$\beta = 0$ when True $\beta = -0.3$					
2SLS	t-Test	.018	.400	1.000	.042	.861	1.000
	False Positives	.012	.000	.000	.007	.000	.000
LIML	CLR Test	.406	.837	1.000	.685	.994	1.000
		Reject H_0:$\beta = 0$ when True $\beta = 0.3$					
2SLS	t-Test	.847	.879	.998	.982	.991	1.000
LIML	CLR Test	.195	.452	.989	.337	.788	1.000
		Median Bias					
2SLS		.231	.105	.022	.235	.097	.016
LIML		.003	.000	.000	.000	-.001	.000

Note: The table reports the frequency of rejecting the null hypothesis H_0: $\beta = 0$.

As we see in Table 10 the performance of 2SLS and the t-test is remarkably poor in this case. First, the median bias is substantial. It is 0.23 for both K=10 and K=20.

[26]Recall that $C = NR^2/(1-R^2)$ where first-stage population R^2 is $Var(z\pi)/(Var(z\pi)+1)$ in our DGP. The R^2 with 1, 3, 10 and 20 instruments is .0023, .0069, .022 and .044.

Second, the size distortion is substantial. With 10 instruments the 5% two-tailed t-test rejects the true null of $\beta = 0$ at a 38.4% rate, increasing to 60.4% with 20 instruments. Even more striking is that all these rejections occur when $\hat{\beta}_{2SLS}$ is positive.

Third, the t-test has essentially no power to detect true negative effects of substantial magnitude. Rates of rejecting H_0:$\beta = 0$ when true $\beta = -0.30$ are very low, and often occur when $\hat{\beta}$ is positive, so we conclude β is positive when it is actually negative.

Fourth, if $\beta = 0.30$ the t-test rejects H_0:$\beta = 0$ at a 84.7% rate with 10 instruments, and with near certainty with 20, but this merely reflects the extreme bias of the procedure towards concluding β is positive. If a researcher is determined to find a significant effect in the same direction as OLS, then the use of the 2SLS t-test is a good choice, and the use of many instruments even better.

In contrast, the performance of LIML combined with CLR is impressive. As expected, size is close to 5%, but more surprisingly, rejections are evenly balanced between positive and negative $\hat{\beta}$. The CLR test has good power to detect true negative effects: The rate of rejecting H_0:β=0 if true β=-0.30 is 40.6% with 10 instruments, increasing to 68.5% with 20 instruments. Of course, CLR has less power on the positive side.

The second (higher) level of instrument strength we consider in Table 10 is C=61.59 (138.33) for K=10 (20). A first-stage $\hat{F} \geq 10$ is required to have 95% confidence that C is at least this high. So this corresponds to the Staiger-Stock rule of thumb.[27] At this level of instrument strength the bias in 2SLS and size distortion in the t-test are still substantial. With 10 instruments the 5% two-tailed t-test rejects the true null of $\beta = 0$ at a 19.9% rate, increasing to 28.9% with 20 instruments. Again, all these rejections occur when $\hat{\beta}$ is positive. In contrast, CLR test size is near 5%, with an even balance of positive vs. negative rejections. The CLR test also has much better power than the t-test to detect true negative effects. The rate of rejecting H_0:$\beta = 0$ when true $\beta = -0.30$ is 83.7% for CLR vs. only 40% for t.

Finally, we consider the very high level of instrument strength of C=314.9 (1114.0) for K=10 (20). A first stage $\hat{F} \geq 38.54$ (62.3) is required to have 95% confidence that C is at least this high. These are the Stock-Yogo test levels for a maximal size distortion of no more than 10% for the t-test. At this very high level of instrument strength the size of the 5% level t-test drops below 10% (to 8.2%), as expected based on the Stock-Yogo analysis. But almost all rejections still occur when $\hat{\beta} > 0$. As we have seen, the power asymmetry in the t-test vanishes very slowly as instrument strength increases. Notice that at this high level level of instrument strength both the CLR and t-tests detect true effects as large as $\beta = \pm 0.30$ with near certainty.

An important point for applied researchers to be aware of is that the difference between LIML and 2SLS is very systematic. Both estimators are consistent, so they converge as C grows large. But one can show the 2SLS estimator is a weighted average of LIML and OLS, so it always lies in between.[28] Hence 2SLS "puts back" some of the OLS bias. As a result, $\hat{\beta}_{2SLS} - \hat{\beta}_{LIML}$ is almost always the same sign as the bias. For example,

[27] Stock and Yogo (2005) note that the required level of $\hat{F}$ to give 95% confidence that the 2SLS bias is less than 10% of the OLS bias is roughly 11 for all values of $K > 3$, and state "this provides a formal ... testing interpretation of the Staiger-Stock rule of thumb" as 11 is close to 10.

[28] The k-class estimators use $kz\hat{\pi} + (1-k)x$ as the instrument for x. LIML is a member of the k-class where $k = 1/(1 - R^2_{\hat{u},\mathbf{z}}) = 1/(1 - Sargan/N) > 1$. If we define a weight $W = (k-1)X'X/k\hat{X}'X \in (0,1)$ we can write $\hat{\beta}_{2SLS} = (1-W)\hat{\beta}_{LIML} + W\hat{\beta}_{OLS}$. LIML is usually defined as choosing $\hat{\beta}$ to minimize k, which is equivalent to minimizing Sargan. A higher k means the instruments suffer more finite-sample contamination, so LIML "pushes away" from OLS.

in Table 8 there are only 1.5% of runs in the weakest instrument case (C=6.9) where $\hat{\beta}_{2SLS} - \hat{\beta}_{LIML}$ is not the same sign as the OLS bias (positive). These are all extreme outliers where $\hat{\beta}_{2SLS} > \hat{\beta}_{OLS}$. In such cases LIML "blows up" to large positive values. If such behavior is observed one should not abandon LIML and adopt 2SLS, as it means $\hat{\beta}_{2SLS}$ is itself suspect. Such behavior never occurs in the stronger instrument cases.

The fact that $\hat{\beta}_{2SLS} - \hat{\beta}_{LIML}$ is almost always the same sign as the OLS bias means LIML is often closer to the true value than 2SLS. In the 10 instrument case, the LIML estimate is closer to the true value in 72%, 67% and 58% of runs, at the three levels of instrument strength. In the 20 instrument case this increases to 81%, 72% and 59%.

Finally, we consider the JIVE estimator that we describe in Appendix D, as the literature often recommends using JIVE with many instruments. Comparing Tables 10 and 11 we see the size of the two-tailed 5% JIVE t-test is less inflated than the 2SLS t-test. But JIVE suffers from the same power asymmetry as 2SLS: In the $F_{5\%} = 5.19$ and 10 cases the large majority of rejections occur when $\hat{\beta} > 0$, and the JIVE t-test has poor power to detect true negative effects. With 10 instruments in the $F_{5\%} = 5.19$ case it only rejects H_0:$\beta = 0$ in 1.7% of cases when true $\beta = -0.30$, compared to 40.6% for LIML + CLR. In this respect JIVE + t offers only a small improvement over 2SLS + t. When the true β is positive (i.e., the direction of the OLS bias) the JIVE t-test has an inflated rejection rate compared to CLR, but this problem is not nearly as bad as for the t-test – i.e., 19.5% vs. 25.3% vs. 84.7%. This is largely because the median bias in JIVE is much less than in 2SLS.

Table 11. JIVE + t-Test Results with Many Instruments ($\rho = 0.8$) (%)

		10 Instruments			20 Instruments		
$F_{5\%Crit}$		5.19	10.00	38.54	4.61	10.00	62.30
		Reject H_0:$\beta = 0$ when True $\beta = 0$					
JIVE	t-Test	.048	.040	.045	.049	.052	.051
	False Positives	.048	.037	.028	.049	.036	.029
		Reject H_0:$\beta = 0$ when True $\beta = -0.3$					
JIVE	t-Test	.017	.689	1.000	.299	.981	1.000
		Reject H_0:$\beta = 0$ when True $\beta = 0.3$					
JIVE	t-Test	.253	.485	.990	.370	.777	1.000
		Median Bias					
JIVE		-.069	-.025	-.005	-.036	-.013	-.002

Note: The table reports the frequency of rejecting the null hypothesis H_0: $\beta = 0$.

Based on these results, we conclude that no case can be made for using 2SLS t-tests in the over-identified case. It is seriously biased towards finding significant effects in the same direction as OLS. The power asymmetry of the t-test combines with the median bias of 2SLS towards OLS to make this bias very strong in the many instrument case. The use of LIML combined with the CLR test avoids both problems, so we strongly recommend this approach. Finlay and Magnusson (2009) provide a heteroskedasticity robust implementation of the CLR test in Stata, that will also invert the test to form confidence intervals. The use of JIVE also avoids median bias towards OLS, but the JIVE + t combination sacrifices a great deal of power compared to LIML + CLR. In results not reported we find the Fuller estimator behaves similarly to LIML, but slightly worse.

7. CONCLUSION

How strong must instruments be for 2SLS and the associated t-test to exhibit acceptable properties? Staiger-Stock suggested the popular rule of thumb that first-stage $\hat{F}$ should be at least 10 for 2SLS t-tests to give reliable results. And, in the case of a single instrument, Stock-Yogo showed that a first-stage $\hat{F}$ of 16.4 gives high confidence that size inflation in two-tailed 2SLS t-tests is no more than 5%. However, we find 2SLS estimates and t-tests are very poorly behaved in environments characterized by $\hat{F}$ in this 10 to 16.4 range.

The Stock and Yogo (2005) focus on size inflation of two-tailed t-tests masks other key problems. First, 2SLS t-tests have very low power for $\hat{F}$ in the 10 to 20 range deemed acceptable by conventional weak IV tests. Second, *the 2SLS estimator has the unfortunate property that it generates standard errors that are artificially small precisely when it generates estimates that are most shifted in the direction of the OLS bias.* Consequently, nearly all significant 2SLS estimates are severely shifted towards OLS when instruments are weak. Surprisingly, this power asymmetry persists if instruments are quite strong.

One consequence of the association between 2SLS estimates and their standard errors is that 2SLS t-tests have poor power to detect true negative effects when the OLS bias is positive. This is true even if instruments are quite strong by conventional standards. This lack of power is of great practical importance, as it means there is little chance of detecting negative program effects given positive selection on unobservables.

A second consequence is that size distortions in one-tailed t-tests are far greater than two-tailed tests. For example, Lee et al. (2022) show a first-stage $\hat{F}$ of 104.7 is sufficient to eliminate size inflation in two-tailed t-tests. But we find a first-stage $\hat{F}$ in the thousands is required to eliminate size distortions in one-tailed 2SLS t-tests.

Applied researchers rarely use one-tailed tests because they expect two-tailed tests to be symmetric (e.g., a two-tailed 5% test is equivalent to a one-tailed 2.5% test). But that is completely false with 2SLS: Even with quite strong instruments most estimates judged significant by two-tailed 2SLS t-tests are shifted in the direction of the OLS bias, rather then symmetrically distributed around the true value.

The power asymmetry in 2SLS t-tests is important for applied work. Take the classic problem of estimating the effect of education on wages. The usual concern is that unmeasured ability biases the OLS estimate upward. But if the OLS bias is indeed positive, then larger 2SLS estimates of the effect of education on wages will *spuriously* appear more precise. This will naturally bias researchers towards exaggerating the effect of education.

Anderson and Rubin (1949)'s test largely avoids the problems that plague the t-test. AR has correct size regardless of instrument strength. Hence, it is widely recommended by theorists for use in just-identified models with weak instruments. Furthermore, it is the most powerful unbiased test in the single instrument case, and it does not sacrifice power to the t-test when instruments are strong. Importantly, we show the AR test is far less susceptible to the power asymmetry that afflicts the t-test. In particular, AR has far better power to detect negative effects when the OLS bias is positive. Thus, we advise discarding the 2SLS t-test altogether, and using AR even with strong instruments.

The AR test is also simple to implement, via OLS estimation of the reduced form, followed by testing for the significance of the instrument. To illustrate, we present an application to estimating the excess sensitivity of consumption to income using PSID data. This allows us to assess the relative performance of AR and t-tests in a realistic setting where the first-stage $\hat{F}$ is modestly above the threshold of 10. We show that in this context the AR test is clearly superior to the t-test in terms of both power and size.

In over-identified models the size inflation in 2SLS t-tests becomes much more severe, and median bias of the 2SLS estimator towards OLS becomes substantial. The use of multiple instruments also increases the covariance between 2SLS estimates and their standard errors, so the power asymmetry of the t-test is amplified. The t-test has low power to detect true effects that are opposite in sign to the OLS bias, and is seriously biased towards finding significant effects in the same direction as the OLS bias.

The bias and size problems that afflict 2SLS and the t-test in the over-identified case led Angrist and Pischke (2008) to argue that applied researchers should choose their one best instrument. However, the use of multiple instruments can increase the efficiency of estimation considerably. Thus, we argue it is important to use methods that exploit the information in multiple instruments without generating bias and size distortions.

In fact, the limited information maximum likelihood (LIML) estimator of Anderson and Rubin (1949) does not suffer from median bias in the multiple instrument case, and the conditional likelihood ratio (CLR) test has much better power properties than the t-test. The use of LIML combined with CLR allows one to exploit the information in multiple instruments without creating bias and size distortion. Hence, in over-identified models we recommend using LIML and CLR in lieu of 2SLS and the t-test, even when instruments are strong. For heteroskedastic data, we advise using a robust version of CLR in conjunction with either LIML or, to gain efficiency, GMM-CU.

In conclusion, we note that recent papers by Andrews et al. (2019) and Young (2022) have emphasized that 2SLS can suffer from low power and size distortions in environments with heteroskedastic and/or clustered errors, even if conventional F tests appear acceptable. We complement that work by showing how similar problems may arise even in *iid* normal settings when instruments are acceptably strong by conventional standards.

ACKNOWLEDGEMENTS

We thank Isaiah Andrews, Don Andrews, Josh Angrist, Michal Kolesar, Robert Moffitt, Whitney Newey and Peter Phillips, as well as Elie Tamer and two anonymous referees, for valuable comments, and Marcelo Moreira for providing his code for conditional t-tests. This research was supported by ARC grants DP210103319 and CE170100005.

REFERENCES

Altonji, J. G. and A. Siow (1987). Testing the response of consumption to income changes with (noisy) panel data. *The Quarterly Journal of Economics 102*(2), 293–328.

Anderson, T. W. and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Math. Statistics 20*(1), 46–63.

Andrews, D., M. Moreira, and J. Stock (2007). Performance of conditional Wald tests in IV regression with weak instruments. *Journal of Econometrics 139*(1), 116–132.

Andrews, I., J. Stock, and L. Sun (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics 11*, 727–753.

Angrist, J. and M. Kolesár (2021). One instrument to rule them all: The bias and coverage of just-id IV. Technical report, National Bureau of Economic Research.

Angrist, J. and J. S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Attanasio, O. P. and M. Browning (1995). Consumption over the life cycle and over the business cycle. *The American Economic Review*, 1118–1137.

Bound, J., D. Jaeger, and R. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association 90*(430), 443–450.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 1365–1387.

Dufour, J.-M. (2004). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics 36*(4), 767–808.

Finlay, K. and L. Magnusson (2009). Implementing weak-instrument robust tests for a general class of instrumental-variables models. *The Stata Journal 9*(3), 398–421.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics 14*(3), 262–280.

Keane, M. and T. Neal (2022a). Robust inference for the Frisch labor supply elasticity. *UNSW Economics Working Paper 2021-07b*.

Keane, M. P. and T. Neal (2022b). A practical guide to weak instruments. *UNSW Economics Working Paper No. 2021-05c*.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica 70*(5), 1781–1803.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022). Valid t-ratio inference for IV. *American Economic Review 112*(10), 3260–90.

Mariger, R. and K. Shaw (1993). Unanticipated aggregate disturbances and tests of the life-cycle consumption model using panel data. *Review of Econ. and Stat. 75*, p.48–56.

Mikusheva, A. and B. P. Poi (2006). Tests and confidence sets with correct size when instruments are potentially weak. *The Stata Journal 6*(3), 335–347.

Mills, B., M. Moreira, and L. Vilela (2014). Tests based on t-statistics for IV regression with weak instruments. *Journal of Econometrics 182*(2), 351–363.

Moreira, H. and M. Moreira (2019). Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors. *J. of Econ. 213*(2), 398–433.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica 71*(4), 1027–1048.

Moreira, M. J. (2009). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics 152*(2), 131–140.

Mork, K. A. and V. K. Smith (1989). Testing the life-cycle hypothesis with a Norwegian household panel. *Journal of Business & Economic Statistics 7*(3), 287–296.

Nelson, C. R. and R. Startz (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business*, S125–S140.

Phillips, G. D. A. and C. Hale (1977). The bias of instrumental variable estimators of simultaneous equation systems. *International Economic Review 18*(1), 219–228.

Phillips, P. C. (1989). Partially identified econometric models. *Econometric Theory 5*(2), 181–240.

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

Stock, J. and M. Watson (2015). *Introduction to econometrics (3rd global ed.)*. Pearson.

Stock, J., J. Wright, and M. Yogo (2002). A survey of weak instruments & weak identification in generalized method of moments. *J of Bus. & Econ. Stat. 20*(4), 518–529.

Stock, J. and M. Yogo (2005). Testing for weak instruments in linear IV regression. *Identification and inference for econometric models 80*(4.2), 1.

Van de Sijpe, N. and F. Windmeijer (2022). On the power of the conditional likelihood ratio and related tests for weak-instrument robust inference. *Journal of Econometrics*.

Young, A. (2022). Consistency without inference: Instrumental variables in practical application. *European Economic Review*, 104112.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Stat. Assc. 57*(298), 348–368.

APPENDICES

A. ANALYTICAL POWER FUNCTIONS OF THE AR AND T-TESTS

Consider the just-identified *iid*-normal linear IV model of equation (1). The power of both the AR and t-tests depends on the true β, the degree of endogeneity ρ, and λ (= square root of population F). The power of the AR test is simply:

$$Power_{AR}(\beta|\lambda,\rho) = \Phi(\lambda D - z_{1-\alpha/2}) + \Phi(-z_{1-\alpha/2} - \lambda D) \tag{A1}$$

where Φ is the standard normal cdf, $D = \beta/\sqrt{Var(v)}$ where $v = \beta e + u$ is the reduced form error with $Var(v) = 1+2\rho\beta+\beta^2$, and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Below we set $\alpha = 0.05$.

Following the the analysis in Stock and Yogo (2005), Lee et al. (2022) and Angrist and Kolesár (2021), the power of the two-tailed 2SLS t-test is given by the integral:

$$\begin{aligned} &Power_t(\beta|\lambda,\rho) = \\ &\int_{-\infty}^{\infty} \left(\mathbb{I}\{t^2 \geq (1-\rho_0^2) z_{1-\alpha/2}^2\} f(t, D, \lambda, \rho_0) + \mathbb{I}\{t^2 \geq z_{1-\alpha/2}^2\} \right) \phi(t-\lambda) dt \end{aligned} \tag{A2}$$

where ϕ is the standard normal density, ρ_0 is the correlation of the reduced form errors, given by $\rho_0 = corr(\beta e + u, e) = (\rho+\beta)/\sqrt{1+2\rho\beta+\beta^2}$, and:

$$f(t, D, \lambda, \rho_0) = \Phi\left(\frac{a_2 - \lambda D - \rho_0(t-\lambda)}{\sqrt{1-\rho_0^2}} \right) - \Phi\left(\frac{a_1 - \lambda D - \rho_0(t-\lambda)}{\sqrt{1-\rho_0^2}} \right), \tag{A3}$$

$$a_1 = \frac{\rho_0 z_{1-\alpha/2}^2 t - |t| z_{1-\alpha/2} \sqrt{t^2 - (1-\rho_0^2) z_{1-\alpha/2}^2}}{z_{1-\alpha/2}^2 - t^2},$$

$$a_2 = \frac{\rho_0 z_{1-\alpha/2}^2 t + |t| z_{1-\alpha/2} \sqrt{t^2 - (1-\rho_0^2) z_{1-\alpha/2}^2}}{z_{1-\alpha/2}^2 - t^2}.$$

The integral in (A2) must be evaluated numerically.

A.1. Power of the AR vs t-test

We now present an example power comparison between the AR and t-test. Consider the case of C=2.3. We need $\hat{F} \geq 10$ to have 95% confidence that C is at least 2.3, so the Staiger-Stock rule of thumb implies this is an acceptable level of instrument strength. We set $\beta = 0$ and $\rho = 0.5$, so the OLS bias is positive, $E(\hat{\beta}_{OLS}) = 0.5$.

Results are shown in Figure A1. The severe power asymmetry of the t-test is evident. Power is less than 5% over the 0 to -0.70 range. As a one sigma change in x induces roughly a β sigma change in y these would be very large effects in most empirical applications. Notice that t-test size is 3% when true $\beta = 0$. The literature often calls a 5% test that rejects at a lower rate than 5% "conservative," and there is a tendency in the weak IV literature to consider this acceptable. But it seems odd to call the t-test conservative when it has no power to detect a large range of true negative effects.

Figure A1 also shows the AR test is unbiased, has correct size, and has superior power to detect true negative effects. It's power reaches 23% when true β is -0.70, compared to only 5% for the t-test. Notably, the power of both tests is uninspiring, which is why we call for a higher standard of instrument strength than $\hat{F} \geq 10$ in empirical practice.

Figure A1. Power of the T-Test vs. AR-Test when $F = 2.3$ ($\rho = 0.5$)

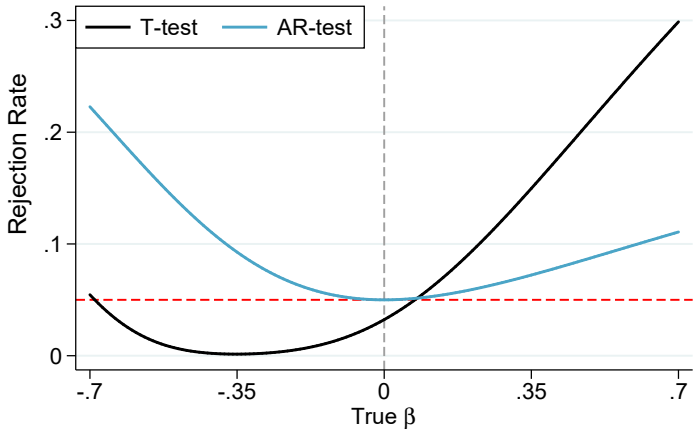


Figure A1 also shows how the t-test power curve rises above AR for positive values of β. As we showed in Section 3, this occurs because 2SLS standard errors are spuriously small in samples where the 2SLS estimate is shifted in the direction of the OLS bias.

One may understand why the power of AR is asymmetric about zero by using (1) to write the reduced form as $y = z\xi + (1 + \beta\rho)u + \beta(1 - \rho^2)^{1/2}\eta$. Assuming $\rho > 0$, as β increases the error variance increases, making the OLS estimate of the reduced form parameter $\xi = \beta\pi$ less precise. The denominator of the regression F-stat also increases, dampening increases in AR generated by increasing β. The AR test properly interprets the increased noise in the reduced form as implying less certainty about significance of $\hat{\beta}$. The 2SLS standard error calculation $Var(\hat{\beta}) = Var(\hat{\beta}_{OLS})/R^2_{x,z}$ ignores this information.

Conversely, the error variance of the reduced form is minimized when $\beta\rho{=}{-}1$, so the power of AR spikes at that point, as the OLS estimate of $\xi = \beta\pi$ is relatively precise. This helps the AR test to have very good power on the negative side if the OLS bias is positive. As $|\beta| \to \infty$ the relationship between x and y becomes deterministic, so the chance of finding a significant relationship between y and z is exactly equal to the chance of finding a significant relationship between x and z. Thus as $|\beta| \to \infty$ the power of AR approaches $P(\hat{F} > 3.84|C)$, the probability of a significant first stage.

B. SIMULATING THE DISTRIBUTION OF THE T-TEST

Following Mills et al. (2014), we begin by defining the covariance matrix of the reduced form errors Ω. The reduced-form equations are $y = z\beta\pi{+}(\beta e{+}u) = z\xi{+}v$ and $x = z\pi{+}e$. Thus we have $\Omega = \begin{pmatrix} \sigma_v^2 & \rho_0\sigma_v\sigma_e \\ \rho_0\sigma_v\sigma_e & \sigma_e^2 \end{pmatrix}$ where $\sigma_v^2 = Var(\beta e + u)$ and $\rho_0 = corr(\beta e + u, e) = corr(v, e)$. Mills et al. (2014) show that:

$$t_{2SLS} = \frac{\beta_{2SLS}}{\sigma_{2SLS}[c_{21}^2 t_{AR}^2 + 2c_{21}c_{22}t_{AR}t_{1'} + c_{22}^2 t_{1'}^2]^{-1/2}} \tag{B1}$$

where t_{AR} is the t-statistic from a regression of y_i on the instrument z_i (i.e. the "t-test version" of the AR test statistic),[29] $t_{1'}$ is the t-statistic of a regression of $x_i - \rho_0\frac{\sigma_e}{\sigma_v}y_i$

[29]Obviously the AR test is equivalent to the squared t-test for significance of the instrument z in the reduced form for y. We denote this by t_{AR} and refer to it as the "t-test version" of the AR test. It is

on z_i, and σ_{2SLS} is the standard error of the 2SLS regression. We also have $c_{11} = \sigma_v$, $c_{21} = \rho_0\sigma_e$, and $c_{22} = \sigma_e\sqrt{1-\rho_0^2}$. Furthermore, β_{2SLS} is given by:

$$\beta_{2SLS} = \frac{c_{11}c_{21}t_{AR}^2 + c_{11}c_{22}t_{AR}t_{1'}}{c_{21}^2t_{AR}^2 + 2c_{21}c_{22}t_{AR}t_{1'} + c_{22}^2t_{1'}^2} \tag{B2}$$

Importantly, the t-test version of the AR test t_{AR} and the modified first-stage t-statistic $t_{1'}$ are constructed to be independent. To see this, note that:

$$\begin{pmatrix} y_i \\ x_i - \rho_0\frac{\sigma_e}{\sigma_v}y_i \end{pmatrix} \sim N\left[\begin{pmatrix} z_i\pi\beta \\ z\pi(1-\rho_0\frac{\sigma_e}{\sigma_v}\beta) \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_e^2(1-\rho_0^2) \end{pmatrix}\right]$$

Thus y_i and $x_i - \rho_0\frac{\sigma_e}{\sigma_v}y_i$ are independent, which implies that t_{AR} and $t_{1'}$ are independent.[30] This allows us to simulate the distribution of t_{AR} while holding $t_{1'}$ fixed.

It is possible to use these equations to simulate the distribution of the 2SLS t-test under the null that H_0:β=0, and conditional on the strength of the instruments (captured by $t_{1'}$) and the covariance of the reduced form errors (Ω), using the following procedure:

1 Draw a simulated value of t_{AR} from the $N(0,1)$ distribution holding $t_{1'}$ fixed.[31]
2 Calculate β_{2SLS} as above but using the simulated value of t_{AR}.
3 Re-estimate σ_{2SLS} using the value of β_{2SLS} from step 2.
4 Calculate t_{2SLS} using the values from the first to third step.
5 Repeat Steps 1-4 N times.
6 To obtain (simulated) critical values for α-level one-sided conditional t-tests, calculate the α and $1-\alpha$ percentiles of the N simulated values of t_{2SLS}.

The 2.5 and 97.5 percentiles from step 6 can be used as critical values for a 5%-level two-sided conditional t-test of the H_0:β=0. We call this an ACT test in the text.

The above procedure assumes that Ω is known, but $\hat{\Omega}$ must be used in practice. This is not important in theory (or practice given large samples), as the covariance structure of the reduced from errors can be consistently estimated without knowing true β or ρ.[32] The procedure can also be made robust to non-normal disturbances due to heteroskedasticity, autocorrelation (in the case of time series or panel data), or clustering by adjusting the standard errors in the reduced form regressions used in the calculation of t_{AR} and $t_{1'}$.

Construction of Table 5 in the main text requires running a simulation within a simulation. For each dataset drawn from our DGP, we obtain estimates of instrument strength and the error covariance structure. Conditional on those estimates, we simulate the conditional distribution of the t-statistic using the above algorithm.

obvious that t_{AR} is approximately standard normal (in large samples) regardless of the weakness of the instrument, as it is simply a t-test from an OLS regression.

[30] Note that t_{AR} is obtained from a projection of y_i on z_i, and $t_{1'}$ is obtained from a projection of $x_i - \rho_0\frac{\sigma_e}{\sigma_v}y_i$ on z_i. Since y_i and $x_i - \rho_0\frac{\sigma_e}{\sigma_v}y_i$ are independent, the independence of the two objects is preserved by these projections. For this reason, $t_{AR}|t_{1'} \sim t_{AR} \sim N(0,1)$.

[31] In the case of multiple instruments ($k \geq 2$), one would draw t_{AR} from a $k \times 1$ vector of $N(0,1)$ draws.

[32] ACT test results in this article are based on our own Stata code. We are thankful to Marcelo Moreira for providing his Matlab code that served as a guide. The original Matlab code used a simulation design in which ρ implicitly changed as β varied from zero, in such a way as to ensure ρ_0 was held constant across scenarios. Our designs hold ρ fixed as we vary β, which leads to very different results. We prefer the fixed ρ design for the reasons explained in Van de Sijpe and Windmeijer (2022). The difference in simulation design also explains to some extent why the results for the tF test in Section 5.2 may look different to those in Lee et al. (2022), as they also allow ρ to change as β changes.

C. THE LIKELIHOOD RATIO (LR) TEST

The likelihood ratio (LR) test developed by Anderson and Rubin (1949) can be used to test the null hypothesis H_0:$\beta = \beta_0$ when the parameter β_0 is estimated by IV regression. Similar to the AR test, the LR test can also be inverted to form confidence intervals. Here we explain how the LR test is constructed. We focus on H_0:$\beta = 0$ to minimize notation. The LR test is based on the reduced form system of two equations:[33]

$$\begin{aligned} y &= \boldsymbol{z}(\beta\boldsymbol{\pi}) + v \\ x &= \boldsymbol{z}\boldsymbol{\pi} + e \end{aligned} \tag{C1}$$

where the error terms e and v have variance-covariance matrix Ω.

Equation (C1) can be estimated as a SUR system. The LR test compares the likelihood of two alternative SUR systems: First, fix β at the LIML estimate, $\hat{\beta}_L$, and estimate the $k \times 1$ parameter vector $\boldsymbol{\pi}$, obtaining $\hat{\boldsymbol{\pi}}_L$. Second, fix $\beta = 0$ and estimate $\boldsymbol{\pi}$, obtaining $\hat{\boldsymbol{\pi}}_0$.

Note that the first SUR constrains the coefficients on z in the y equation to be $\hat{\beta}_L$ times their coefficients in the x equation. The second SUR drops z from the y equation completely, introducing one additional constraint. Letting $\ell(\hat{\beta}_L, \hat{\pi}_L)$ and $\ell(0, \hat{\pi}_0)$ denote the log-likelihoods of the two SUR systems, the LR test for H_0:$\beta = 0$ is simply $LR = 2(\ell(\hat{\beta}_L, \hat{\pi}_L) - \ell(0, \hat{\pi}_0))$. It is distributed $\chi^2(1)$ under the null.

The likelihood for a regression with normal errors takes a simple form. Let $V_i = \left(\begin{smallmatrix} v_i \\ e_i \end{smallmatrix}\right)$ denote the error vector for observation i, with $V_i \sim N(0, \Omega)$. The likelihood is the bivariate normal density of the errors for the N observations:

$$L(\beta, \pi|\Omega) = \frac{1}{(2\pi)^N}(|\Omega|)^{-N/2} exp\left[-\frac{1}{2}\sum_{i=1}^{N} V_i' \Omega^{-1} V_i\right] \tag{C2}$$

Now we form the sample likelihood for each of the two SUR models. The vector of sample residuals for observation i is $\hat{V}_i = \left(\begin{smallmatrix} y_i - \mathbf{z}_i\hat{\beta}\hat{\boldsymbol{\pi}} \\ x_i - \mathbf{z}_i\hat{\boldsymbol{\pi}} \end{smallmatrix}\right)$, where $(\hat{\beta}, \hat{\boldsymbol{\pi}})$ is either $(\hat{\beta}_L, \hat{\boldsymbol{\pi}}_L)$ or $(0, \hat{\boldsymbol{\pi}}_0)$. Let $\hat{\Omega}$ denote the sample variance-covariance matrix of the residuals, which is $\hat{\Omega}_L$ or $\hat{\Omega}_0$.

Notice that if we plug the sample residuals $\hat{V}_i$ and sample covariance matrix $\hat{\Omega}$ into (C2) the sum of squared residuals term in square brackets simplifies tremendously. The normalization by $\hat{\Omega}$ renders the residuals independent standard normal, so the sum of squared residuals is simply $2N$. Thus the sample likelihood is:

$$L(\hat{\beta}, \hat{\pi}|\hat{\Omega}) = \frac{1}{(2\pi)^N}(|\hat{\Omega}|)^{-1/2} exp\left[-N\right] \tag{C3}$$

Taking logs and cancelling like terms, the likelihood ratio simplifies to:

$$LR = 2\left(\ell(\hat{\beta}_L, \hat{\pi}_L|\hat{\Omega}_L) - \ell(0, \hat{\boldsymbol{\pi}}_{\mathbf{0}}|\hat{\Omega}_0)\right) = N\left(log(|\hat{\Omega}_0|) - (log(|\hat{\Omega}_L|)\right) \tag{C4}$$

Thus, the LR test simply compares the estimated variance-covariance matrix of the residuals in two versions of the reduced form system, with β set to either $\hat{\beta}_L$ or 0. Intuitively, if the residual variance increases significantly when we constrain $\beta = 0$, we conclude that the estimate of β is significant.[34]

As we saw in Section 6.2 the LR test suffers a modest size distortion when instruments

[33] As with other measures of regression fit, such as R^2, calculating the likelihood for the structural equation $y = \beta x + u$ makes little sense as the likelihood will often be worse under the optimised coefficient estimates than the null hypothesis.

[34] Notice that $|\Omega| = \sigma_e^2\sigma_v^2 - 2Cov(e, v)$. So the overall residual variance of the reduced form system is increasing in σ_e^2 and σ_v^2 and decreasing in the covariance.

are weak, although far less so than 2SLS. Moreira (2003) developed a conditional LR test that adjusts critical values based on instrument strength to eliminate the size distortion. Similar to the conditional t-tests discussed in Section 5 and Appendix B, the idea is to simulate the distribution of the LR test, which depends on instrument strength because the test is not pivotal. To implement this idea, he assumes that Ω is known. Then, plugging the sample residuals $\hat{V}_i$ into (C2) we obtain:

$$L(\hat{\beta},\hat{\pi}|\Omega)=\frac{1}{(2\pi)^N}(|\Omega|)^{-N/2}exp\left[-\frac{1}{2}\sum_{i=1}^{N}\hat{V}_i'\Omega^{-1}\hat{V}_i\right] \tag{C5}$$

Let $\hat{V}_{Li}$ and $\hat{V}_{0i}$ denote the residuals from the SUR models with $(\hat{\beta},\hat{\pi})$ equal to $(\hat{\beta}_L,\hat{\pi}_L)$ and $(0,\hat{\pi}_0)$ respectively. Plug the $\hat{V}_{Li}$ and $\hat{V}_{0i}$ into (C5) to form the likelihoods of the two models. Taking logs and cancelling like terms, this version of the LR test simplifies to:

$$LR_0=2\left(\ell(\hat{\beta}_L,\hat{\boldsymbol{\pi}}_L|\Omega)-\ell(0,\hat{\boldsymbol{\pi}}_0|\Omega)\right)=\left[\sum_{i=1}^{N}\hat{V}_{0i}'\Omega^{-1}\hat{V}_{0i}\right]-\left[\sum_{i=1}^{N}\hat{V}_{Li}'\Omega^{-1}\hat{V}_{Li}\right] \tag{C6}$$

In practice, Ω must be estimated. Moreira (2003) proposes using residuals from estimating the two reduced form equations in (C1) separately by OLS to form $\hat{\Omega}$. When $\hat{\Omega}$ is used in (C6) it is an approximation to the true likelihood ratio statistic in (C4).[35]

Moreira (2003) shows that the approximate LR statistic LR_0 can be written:

$$LR_0=\frac{1}{2}\left(S'S-T'T+\sqrt{(S'S+T'T)^2+4[(S'T)^2-(S'S)\cdot(T'T)]}\right) \tag{C7}$$

where $S=(Z'Z)^{-1/2}Z'Y/\sigma_v$ and $T=(Z'Z)^{-1/2}Z'(X-\rho_0\frac{\sigma_e}{\sigma_v}Y)/\sigma_e\sqrt{1-\rho_0^2}$, and where we have stacked the N observations on y, x and z to form Y and X vectors that are $N\times 1$ and a Z matrix that is $N\times K$. But in the single instrument (K=1) case S is simply the t-test from regressing y on z, that in Appendix B we denoted t_{AR}, and T is the t-test from regressing $x-\rho_0\frac{\sigma_e}{\sigma_v}y$ on z, that we denoted $t_{1'}$.[36] In the over-identified case ($k>1$) $S'S$ is the AR test and $T'T$ is an F-test for instrument strength (both scalars).

In the exactly identified case one can replace S and T in (C7) with the scalars t_{AR} and $t_{1'}$ and it simplifies down to the AR statistic $S'S$ as the term in square brackets vanishes. This shows that the AR and LR statistics are equal in the exactly identified case. And the fact that neither test depends on T shows these tests are pivotal.[37]

When $K>1$ the LR statistic is no longer pivotal as it depends on instrument strength. Equation (C7) no longer simplifies to AR because $S'T$ is no longer equal to $t_{AR}\cdot t_{1'}$ when S and T are vectors. As we discussed in Section 6.1, increasing K mechanically drives up AR ($=S'S$) – which is the NR^2 from regressing y on $\mathbf{z}$ – for the usual reason that there exists sample covariance between each regressor z_k and the errors u. LR corrects for this by subtracting off the Sargan statistic – the NR^2 from regressing $\hat{u}$ on $\mathbf{z}$ – from

[35]Moreira (2003) claims the approximation is very accurate, and in our simulations we have not found any evidence to the contrary.

[36]Recall from Appendix B that y and $x-\rho_0\frac{\sigma_e}{\sigma_v}y$ are independent, so S and T are independent.

[37]S is simply $N(0,1)$ under the null hypothesis no matter the strength of the instrument. To see this simply, consider the reduced form for y when $K=1$: $y=\beta\hat{\pi}z+e$. The value of $\hat{\pi}$, so long as it is not zero, has no influence on how well the regression can fit the data, as $\hat{\beta}$ can adjust freely. In contrast, if K=2 the we have $y=\beta(z_1\hat{\pi}_1+z_2\hat{\pi}_2)+e$. Now, the particular values of $\hat{\pi}_1$ and $\hat{\pi}_2$ do matter in terms of how well the regression can fit the data.

AR, causing the two statistics to diverge. The residuals $\hat{u} = y - \hat{\beta}_L x$ depend on the LIML estimate of β, so the LR test is not pivotal, and its distribution depends on instrument strength as measured by T. This introduces a size distortion to the LR test.

Moreira (2003) developed a conditional version of the LR test that adjusts the critical value as function of instrument strength to correct the size distortion. Given that the vectors S and T are independent, it is possible to use S to simulate the distribution of the LR test under the null that H_0:β=0 and H_1:$\beta \neq 0$, conditional on the strength of the instruments (i.e. holding T fixed across draws of S) and the covariance of the reduced form errors (Ω). He proposes the following procedure:

1 Draw a simulated S as a $k \times 1$ vector of $N(0,1)$ draws, holding T fixed.
2 Calculate LR as in (C7) but using the simulated vector S.
3 Repeat Steps 1-2 J times.
4 To obtain (simulated) critical values for the α-level two-sided Likelihood Ratio test, calculate the α percentiles of the J simulated values of LR.

As in the simulation of the t-test, the above procedure assumes that Ω is known, yet we use $\hat{\Omega}$ in practice. Andrews et al. (2007) provide a method to compute the p-values of the CLR test using numerical integration, which offers a significant computational advantage over simulating the critical values. The current user-written commands in Stata that implement the CLR test, *condivreg* and *weakiv*, both provide p-values using the integration approach. Furthermore, the CLR procedure can be made robust to non-normal disturbances due to heteroskedasticity, autocorrelation (in the case of time series or panel data), or clustering by adjusting the estimate of the reduced form covariance matrix $\hat{\Omega}$. Details can be found in Finlay and Magnusson (2009).

D. THE JIVE ESTIMATOR

2SLS can be interpreted as IV using $z_i\hat{\pi}$ as the instrument for x_i, where $\hat{\pi}$ is obtained from OLS regression of x on z. Obviously $\hat{\pi}$ tends to be greater in samples where $\widehat{cov}(z,e)$ is greater, and this has an unfortunate consequence: For an individual observation i we have that $cov(z_i\hat{\pi}, e_i) > 0$, because a *ceteris paribus* increase in $z_i e_i$ drives up $\hat{\pi}$. If $\rho > 0$ this means $cov(z_i\hat{\pi}, u_i) > 0$, so the instrument is positively correlated with the structural error, which biases the 2SLS median towards OLS.[38]

Phillips and Hale (1977) noted this phenomenon, and suggested an alternative IV estimator using $z_i\hat{\pi}_{-i}$ as the instrument for x_i, where $\hat{\pi}_{-i}$ is obtained from OLS regression of x on z *excluding* observation i. This approach, later called "jackknife IV" (JIVE), breaks the correlation between $z_i\hat{\pi}$ and u_i.

We have emphasized the problem that 2SLS is much more likely to judge estimates significant if they are shifted in the direct of the OLS bias. In Section 6.3 we show that JIVE suffers from the same problem. There exists a strong negative association between $se(\hat{\beta}_{JIVE})$ and $\hat{\beta}_{JIVE}$ imparts positive $\hat{\beta}_{JIVE}$ estimates with spuriously high precision.

In fact, in Keane and Neal (2022b) we show that JIVE can perform worse than 2SLS in some contexts, because the alternative instrument $z_i\hat{\pi}_{-i}$ has a smaller correlation with x than $z_i\hat{\pi}$, making the weak instrument problem worse. This has especially dire consequences if the instrument z is weak to begin with.

[38]The covariance of $z_i\hat{\pi}$ and u_i is of order $1/N$, as the influence of observation i on $\hat{\pi}$ vanishes as N grows large, but in finite samples it contributes to bias in the 2SLS median.

Online Appendices to Accompany:

"Instrument Strength in IV Estimation and Inference: A Guide to Theory and Practice"

Michael Keane and Timothy Neal

Date: November 30, 2022

MULTIPLE INSTRUMENTS: RESULTS WITH LOWER LEVELS OF ENDOGENEITY

In this Online Appendix we examine the properties of the 2SLS and LIML estimators and associated tests given lower levels of endogeneity (as defined by ρ) than in the main text. All simulation results presented here are based on the DGP outlined in (1) of the main article. Other than different values set for ρ, they are comparable to those found in Table 8 in the text. We begin with Table 1, which compares the performance of the 2SLS estimator combined with its t-test with the LIML estimator combined with the CLR test of Moreira (2003). We start by setting $\rho = 0.5$, which represents a moderate level of endogeneity, and compare this case with the high level of endogeneity ($\rho = 0.8$) that was used in Table 8 in the main article.

Many of the conclusions that were drawn in Section 6.2 of the main text carry over to this scenario of more moderate endogeneity. The median bias of 2SLS remains substantial but is less severe than in the $\rho = 0.8$ case. Again, the LIML estimator removes all of the median bias. The size distortion of the t-test becomes less severe, due to the fact that the association between the 2SLS estimate of β and its standard error becomes somewhat weaker. Yet the power asymmetry problem remains serious. At weaker levels of instrument strength, such as $C = 6.9$ or 13, the t-test has almost no power to detect true negative effects in stark comparison to the CLR test using LIML. Even in the much stronger case of $C = 40.9$ the 2SLS t-test has notably inferior power to the CLR test. This combination of median bias and power asymmetry biases the t-test towards finding effects in the direction of the OLS bias (which is positive as $\rho > 0$).

Table 1. 2SLS vs. LIML: Size and Power with K=3 ($\rho = 0.5$) (%)

C		6.90	13.01	40.91	110.55	360.26
$F_{5\%Crit}$		6.93	10.00	22.30	50.00	142.50
		Reject H_0:$\beta = 0$ when True $\beta = 0$				
2SLS	t-Test	.083	.074	.057	.055	.052
	False Positives	.083	.074	.054	.042	.033
LIML	CLR Test	.051	.052	.049	.049	.050
	False Positives	.029	.025	.023	.023	.024
		Reject H_0:$\beta = 0$ when True $\beta = -0.3$				
2SLS	t-Test	.011	.048	.427	.917	1.000
	False Positives	.004	.001	.000	.000	.000
LIML	CLR Test	.129	.210	.566	.941	1.000
	False Positives	.008	.001	.000	.000	.000
		Reject H_0:$\beta = 0$ when True $\beta = 0.3$				
2SLS	t-Test	.330	.367	.555	.849	.999
LIML	CLR Test	.095	.142	.357	.758	.998
		Median Bias				
2SLS		.126	.071	.022	.008	.002
LIML		.016	-.001	-.001	-.001	.000

Note: The table reports the frequency of rejecting the null hypothesis H_0: $\beta = 0$.

Table 2 reports the same set of results with $\rho = 0.2$, which represents a low level of endogeneity. In this scenario, we would expect many of the problems that afflict the 2SLS t-test, such as power asymmetry and median bias, to significantly improve. Remarkably, the table shows that the power asymmetry of the t-test remains significant. The size distortion of the t-test becomes negative, as the rejection rate falls below 5% when $\beta = 0$. Power is much greater in the direction of the OLS bias when $C = 6.9$ or 13. Only when instruments become quite strong, such as $C = 111(F_{5\%Crit} = 50)$, does the t-test power asymmetry become negligible. At that point the behavior of 2SLS and the t-test is very similar to that of LIML and the CLR-test.

Table 2. 2SLS vs. LIML: Size and Power with K=3 ($\rho = 0.2$) (%)

C		6.90	13.01	40.91	110.55	360.26
$F_{5\%Crit}$		6.93	10.00	22.30	50.00	142.50
		Reject H_0:$\beta = 0$ when True $\beta = 0$				
2SLS	t-Test	.022	.030	.041	.048	.049
	False Positives	.022	.027	.029	.029	.027
LIML	CLR Test	.052	.051	.050	.049	.050
	False Positives	.027	.025	.024	.023	.024
		Reject H_0:$\beta = 0$ when True $\beta = -0.3$				
2SLS	t-Test	.059	.138	.475	.892	1.000
	False Positives	.000	.000	.000	.000	.000
LIML	CLR Test	.111	.178	.479	.889	1.000
	False Positives	.007	.002	.000	.000	.000
		Reject H_0:$\beta = 0$ when True $\beta = 0.3$				
2SLS	t-Test	.194	.271	.525	.863	1.000
LIML	CLR Test	.100	.150	.398	.810	.999
		Median Bias				
2SLS		.050	.028	.008	.003	.001
LIML		.007	-.001	-.001	-.001	-.001

Note: The table reports the frequency of rejecting the null hypothesis H_0: $\beta = 0$.

Importantly, Table 2 also reveals that power is poor when ρ=0.2 unless instruments are quite strong, regardless of whether the t-test or CLR-test is used. This highlights how the t-test derives much of its power from finite-sample correlation between the instrument and the structural error, which generates spurious correlation between x and z. This source of power is limited when endogeneity is weak.

It is also worth considering the bottom right panel of Figure 9 in the main article, which shows the rate at which OLS estimates are closer to the truth than 2SLS estimates, as a function of C and ρ. When $\rho = 0.2$, instruments must be quite strong for 2SLS to even outperform the inconsistent OLS estimator of β. Accordingly, while the power asymmetry and median bias of the 2SLS t-test does get weaker as ρ shrinks, other severe problems with estimation and inference present themselves.

These results suggest that the 2SLS t-test is of dubious use for empirical research even at modest levels of endogeneity ρ. Moreover, despite the improvements offered by LIML and the CLR-test, there is no substitute for strong instruments.

Lastly, Table 3 compares the performance of 2SLS and LIML in cases with many instruments (K=10 or 20). It is comparable with Table 10 in the main text, but repeats the analysis with $\rho = 0.5$, a more moderate level of endogeneity. All of the main conclusions drawn from the $\rho = 0.8$ case carry over here. The 2SLS median bias is severe unless instruments are very strong, as is the size distortion of the t-test. With weaker instruments, such as $C = 23$ with K=10 and $C = 46$ with K=20, the CLR-test used in combination with the LIML estimator has significant advantages. It has correct size, and its power is much stronger in the direction away from the OLS bias. Even when instruments are very strong, and the tests have perfect power in both directions, the LIML and CLR-test combination still performs much better in the balance of rejections when $\beta = 0$, and also has the advantage in terms of size and median bias.

Table 3. Results with Many Instruments ($\rho = 0.5$) (%)

		10 Instruments			20 Instruments		
C		23.00	61.59	314.90	46.00	138.33	1114.00
$F_{5\%Crit}$		5.19	10.00	38.54	4.61	10.00	62.30
		Reject H_0:$\beta = 0$ when True $\beta = 0$					
2SLS	t-Test	.170	.103	.063	.273	.140	.067
	False Positives	.170	.101	.054	.273	.139	.056
LIML	CLR Test	.053	.049	.048	.056	.053	.053
	False Positives	.028	.026	.025	.027	.027	.027
		Reject H_0:$\beta = 0$ when True $\beta = -0.3$					
2SLS	t-Test	.097	.531	1.000	.228	.906	1.000
	False Positives	.002	.000	.000	.000	.000	.000
LIML	CLR Test	.295	.711	1.000	.521	.967	1.000
		Reject H_0:$\beta = 0$ when True $\beta = 0.3$					
2SLS	t-Test	.708	.826	.999	.929	.983	1.000
LIML	CLR Test	.192	.476	.994	.336	.814	1.000
		Median Bias					
2SLS		.231	.105	.022	.235	.097	.016
LIML		.003	.000	.000	.000	-.001	.000

Note: The table reports the frequency of rejecting the null hypothesis H_0: $\beta = 0$.