A Practical Guide to Weak Instruments

Michael Keane[†] and Timothy Neal[†]

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

April 11, 2022

Abstract We provide a simple survey of the literature on weak instruments, aimed at giving practical advice to applied researchers. It is well-known that 2SLS has poor properties if instruments are exogenous but "weak." We clarify these properties, explain weak instrument tests, and examine how behavior of 2SLS depends on instrument strength. A common standard for "strong" instruments is a first-stage F-statistic of at least 10. But 2SLS has some poor properties in that context: It has low power, and the 2SLS standard error estimate tends to be artificially small in samples where the 2SLS parameter estimate is most contaminated by the OLS bias. This causes t-tests to give very misleading results. Surprisingly, this problem persists even if the first-stage F is in the thousands. Robust tests like Anderson-Rubin greatly alleviate these problems, and should be used in lieu of the t-test even with strong instruments. In many realistic settings a first-stage F well above 10 may be necessary to give high confidence that 2SLS will outperform OLS. For example, in the archetypal application of estimating returns to education, we argue one needs F of at least 50.

JEL: *C12, C26, C36*

Keywords: *Instrumental variables, weak instruments, 2SLS, endogeneity, F-test, size distortion, Anderson-Rubin test, conditional t-test, conditional LR test, Fuller, JIVE*

1. INTRODUCTION

The past 30 years have seen an explosion of applied work using instrumental variable (IV) methods to deal with endogeneity problems. But since Bound et al. (1995), applied economists have become acutely aware that 2SLS estimators have poor properties when instruments are exogenous but "weak" – meaning they are weakly correlated with the endogenous variable. In particular, if instruments are only marginally significant in the first stage of 2SLS, it is now well understood that both the estimates and their standard errors can be very misleading – even in very large samples.

The first goal of this paper is to explain the complex literature on weak instruments in a way accessible to applied researchers. We explain weak instrument tests, weak instrument robust inference and alternatives to 2SLS. Our main contribution is to highlight some important problems with 2SLS that weak instrument tests gloss over. The behavior of 2SLS t-tests is especially problematic, even in large samples with instruments considered "strong" by conventional standards. This leads us to advocate a higher standard of instrument strength, and wider use of robust tests, in applications of IV. We argue the Anderson and Rubin (1949) test should be widely adopted in lieu of the t-test.

Concern with the poor properties of 2SLS in the weak instrument context led Staiger and Stock (1997) to advocate a higher standard of instrument relevance in the first stage of 2SLS. That is, to be confident the estimator is well-behaved, we should require instrument significance at a level *higher* than 5% in the first stage. They find if the first-stage F is greater than 10 – corresponding to a t of 3.16 (p = .0008), in the one endogenous variable, one instrument case – then 2SLS two-tailed t-tests of the hypothesis $H_0 : \beta = 0$ will reject the null at a rate that is not "too far" from the correct 5% rate.

This $F > 10$ advice has been widely adopted in practice and presented in textbooks. For example, Stock and Watson (2015, p.490) say: "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) proposed critical values for F based on the maximal size distortions in 2SLS hypothesis tests one is willing to tolerate. They find $F > 16.4$ ensures (with 95% confidence) that a two-tailed 5% t-test will reject at a 10% rate or less. A recent paper by Lee et al. (2020) shows one needs $F > 104.7$ for two-tailed 2SLS t-tests to have correct rejection rates – a much higher standard than $F > 10$.

Unfortunately, these results are not well understood by applied researchers. They are derived using non-standard asymptotic theory, or complex small sample approximations, that are unfamiliar to most applied economists. This makes sorting through the diverse advice on acceptable first stage F levels a daunting task. We seek to explain weak instrument tests in a way that is accessible to applied researchers familiar with basic statistics. To make this possible, we focus primarily on the case of a single endogenous variable and a single instrument, as is very common in applied practice – see Andrews et al. (2019).

Since Stock and Yogo (2005), the weak instrument literature has been heavily focused on assessing size distortions in 2SLS t-tests.[1] We argue this has caused the literature to gloss over *other* important problematic properties of 2SLS that persist even when instruments are "strong" according to Stock-Yogo tests, and even in large samples.

In particular, our analysis shows that the behavior of the 2SLS estimator in environments characterized by first-stage F thresholds of 10 or 16.4 (or even higher) is extremely poor. The maximal size distortion of two-tailed 2SLS t-tests is modest, just as Stock-Yogo predict, but this masks far more fundamental problems: 2SLS has very low power, combined with *a very unfortunate tendency to generate artificially low standard errors precisely when it generates estimates most contaminated by endogeneity.*

This association between 2SLS estimates and their standard errors, which persists even if instruments are quite "strong" and samples are large, has two important consequences: First, 2SLS t-tests are much more likely to reject the null $H_0 : \beta = 0$ in samples where the 2SLS estimate is shifted in the direction of the OLS bias. Second, 2SLS t-tests have little power to reject the null if the true β is opposite in sign to the OLS bias.

The practical import of these two facts is serious: In an archetypal application of IV, one seeks to test if a policy intervention has a positive effect on an outcome, but a confound arises because those who receive the intervention tend to be positively selected on unobservables. In such a context, even if instruments are moderately strong by conventional standards, 2SLS will have spuriously inflated power to find false positive effects, and little power to detect true negative effects.

If the first-stage F meets the 105 threshold suggested by Lee et al. (2020) then 2SLS does exhibit nice properties in terms of both two-tailed t-test size and power. But the association between 2SLS estimates and their standard errors still has an important influence that distorts t-test results. 2SLS t-tests are much more likely to reject the null in samples where the 2SLS estimate is shifted in the direction of the OLS bias, and this problem persists unless until the first-stage F is in the thousands.

We go on to examine whether the use of "weak instrument robust" tests like Anderson-Rubin or the "conditional t-tests" of Mills et al. (2014) result in more reliable inferences. We find that these approaches greatly alleviate the problems we identify *provided* the

[1] Stock and Yogo (2005) also considered bias relative to OLS. But this criterion can only be assessed given over-identification of degree 2, which is less common in practice than exact identification.

instruments are strong enough that 2SLS has acceptable power in the first place, which in practice requires a first-stage F threshold well above the conventional level of 10.

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 seems clear the AR test should be widely adopted in applied work.

We then examine performance of the main alternative estimators to 2SLS: the Fuller (1977) estimator, JIVE and the unbiased estimator of Andrews and Armstrong (2017). We find the Fuller and unbiased estimators do offer improvements, but their performance cannot be judged adequate unless the first-stage F threshold is well above 10.

Applied researchers face a choice between 2SLS and OLS, so it is interesting to assess their relative performance. Our simulations suggest first-stage F must be well above 10 to have high confidence 2SLS will outperform OLS, unless endogeneity is quite severe. So unless a higher standard can be met, OLS combined with a serious attempt to control for sources of endogeneity may be a superior research strategy to reliance on IV.

Finally, we examine the over-identified case. More instruments increase efficiency but make 2SLS t-tests more misleading, as the covariance between 2SLS estimates and their standard errors gets stronger. This makes robust inference even more important.

In summary, we find 2SLS performs very poorly when the first-stage F is toward the low end of the 10+ range deemed acceptable by current practice. We argue that a higher threshold should be adopted. Even then, it is essential to use robust tests, like AR or the "conditional t-tests" of Mills et al. (2014) unless first-stage F is in the thousands.

2. SOME BACKGROUND ON THE WEAK INSTRUMENT PROBLEM

To clarify the weak instrument problem we focus mainly on the case of one endogenous variable and one instrument, while assuming *iid* sampling. Issues of heteroskedastcity and multiple instruments are discussed later. We also abstract from exogenous covariates, as their inclusion complicates notation without changing anything of substance. Consider a structural equation where outcome y for person i is regressed on endogenous variable x:

$$y_i = x_i\beta + u_i, \text{ where } cov(x_i, u_i) \neq 0$$

The first-stage regression of x on the exogenous instrument z is:

$$x_i = z_i\pi + e_i, \text{ where } cov(z,u) = 0,\ cov(z,e) = 0,\ \pi \neq 0.$$

The regressor x is endogenous if $cov(e,u) \neq 0$, and the instrument z is valid if $cov(z,u) = 0$ and $\pi \neq 0$. It will often be convenient to assume $\pi > 0$, which can always be achieved by normalizing z. It is useful to decompose the error term e in the first stage into parts that are correlated and uncorrelated with the error u in the structural equation:

$$e_i = \rho u_i + \eta_i \quad \text{where} \quad cov\,(\eta, u) = 0,\ cov\,(z, \eta) = 0 \tag{1}$$

Here ρ controls the severity of the endogeneity problem, and x is exogenous if ρ=0.

The 2SLS estimator of β takes the following form, where ˆ denotes a sample value:

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

Clearly 2SLS is consistent: As $N \to \infty$ the sample covariance $\widehat{cov}(z,u)$ converges to its true value $cov(z,u) = 0$, and $\widehat{cov}(z,x)$ converges to $\pi\sigma_z^2 > 0$. So $\hat{\beta}_{2SLS}$ converges to the true β. But this result is not very useful in practice: Here we are interested in properties of 2SLS in finite samples – including, as we will see, very large finite samples.

We can substitute (1) into (2), and write $\hat{\beta}_{2SLS} - \beta$ in the instructive form:

$$\hat{\beta}_{2SLS} - \beta = \frac{\widehat{cov}\,(z,u)}{\pi\widehat{var}\,(z) + \widehat{cov}\,(z,e)} = \frac{\widehat{cov}\,(z,u)}{\pi\widehat{var}\,(z) + \widehat{cov}\,(z,\eta) + \rho\widehat{cov}\,(z,u)} \tag{3}$$

As a point of comparison, recall that the analogous expression for OLS is simply $\hat{\beta}_{OLS} - \beta = \widehat{cov}\,(x,u)\,/\widehat{var}(x)$, where the denominator only depends on observed covariates. This renders it far easier to analyze. In particular, under standard assumptions, the bias in OLS is simply $\rho var(u)/var(x)$, so OLS is unbiased in finite samples if $\rho = 0$.

Contrary to the simplicity of OLS, the $\widehat{cov}\,(z,\eta)$ and $\rho\widehat{cov}\,(z,u)$ terms sitting in the denominator of (3) render the finite sample properties of 2SLS quite complicated. In fact, as we explain below, 2SLS has odd finite sample properties that create serious problems if instruments are "weak," but that become irrelevant if instruments are "strong." We focus on the case of a perfectly exogenous instrument z with population covariance $cov(z,u)$=0, abstracting from problems created by violations of instrument exogeneity. Importantly, even small violations create severe asymptotic bias if instruments are weak, but this issue is discussed extensively elsewhere (see Wooldridge 5e 2012 p 521).

A crucial point is that in finite samples even a perfect instrument has some sample covariance with the error in the structural equation. Hence $\widehat{cov}(z,u)$ will be always be non-zero, even if the instrument is in fact exogenous. Similarly, $\widehat{cov}(z,\eta)$ will depart from zero in finite samples. As a result, the strength of the relationship between the instrument z and the endogenous variable x fluctuates from sample to sample, being stronger in samples where $\widehat{cov}(z,u)$ and $\widehat{cov}(z,\eta)$ are the same sign as π.

The fact that the sample covariances $\widehat{cov}(z,u)$ and $\widehat{cov}(z,\eta)$ appear in the denominator of (3) has unpleasant consequences for the finite sample behavior of the 2SLS estimator, even in large samples. We can learn a lot about these properties by carefully studying equation (3).[2] Now we list some key properties and provide a simple intuition for each:

First, the mean and variance of the 2SLS estimator do not exist: As $\widehat{cov}(z,u)$ and $\widehat{cov}(z,\eta)$ are both random variables, there is nothing to prevent finite sample realizations where $\widehat{cov}(z,x) = \pi\widehat{var}(z) + \widehat{cov}(z,\eta) + \rho\widehat{cov}(z,u) \approx 0$, causing $\hat{\beta}_{2SlS}$ to explode. Of course, this means the variance of $\hat{\beta}_{2SLS} - \beta$ doesn't exist either.

Second, the distribution of $\hat{\beta}_{2SLS} - \beta$ will exhibit skewness and fat tails (i.e., it is non-normal), rendering conventional t-tests non-normal and hence unreliable as a guide to inference. To see why, assume $\rho > 0$, so the endogeneity bias is positive, and that $\pi > 0$, so z has a positive *population* covariance with x. And to keep things simple, let us assume $\widehat{cov}(z,x) > 0$ so at least the *sample* covariance of z and x is the same sign as the population covariance - although this is not guaranteed.[3] Given these assumptions, the denominator of (3) is positive, so the *sign* of $\hat{\beta}_{2SLS} - \beta$ is determined by the sign of $\widehat{cov}(z,u)$ in the numerator. But the *magnitude* of $\hat{\beta}_{2SLS} - \beta$ is amplified (attenuated) when

[2] Research on finite sample properties of 2SLS with weak instruments relies on non-standard asymptotics – the local-to-zero asymptotics of Staiger-Stock (1997) or many-instrument asymptotics of Bekker (1994) – or on complex small sample theory; see Phillips (1983) or Rothenberg (1984). But it is surprising how much can be learned just by studying equation (3) using basic statistics. That is our approach here.

[3] We may also draw samples with $\widehat{cov}\,(z,x) \approx 0$. This generates extreme positive or negative outliers for $\hat{\beta}_{2SLS}$, depending on the sign of $\widehat{cov}\,(z,u)$. Also, if $\widehat{cov}\,(z,u)$ and/or $\widehat{cov}\,(z,\eta)$ go negative enough to drive $\widehat{cov}\,(z,x)$ negative, the sign of $\hat{\beta}_{2SLS} - \beta$ flips, so the distribution of $\hat{\beta}_{2SLS} - \beta$ can be bi-modal.

$\widehat{cov}(z,u) < 0$ (>0), which shrinks (inflates) the denominator. Thus, positive estimates of β are reined in while negative estimates are inflated, and the distribution of $\hat{\beta}_{2SLS} - \beta$ is skewed to the left. [We'll see this clearly in Section 4.]

Third, a large positive realization of $\rho\widehat{cov}(z,u)$ generates both an estimate shifted towards OLS and a low standard error - as it causes a large $\widehat{cov}(z,x)$. *Thus 2SLS has the unfortunate property that it gives artificially low standard errors in samples where $\hat{\beta}_{2SLS}$ is most shifted towards OLS.* As we'll see below, this association between 2SLS estimates and their standard errors has very important consequences that appear to have been largely neglected, or at least not adequately explored, in the prior literature.

Fourth, the median of $\hat{\beta}_{2SLS}$ is biased in the direction of OLS if the instrument is "weak." Recall that OLS is biased in a positive direction if $\rho > 0$. To see that 2SLS is biased in the same direction, consider the extreme case where the instrument is completely irrelevant, $\pi = 0$, so that $x_i = \eta_i + \rho u_i$. Then from (3) we have:

$$\hat{\beta}_{2SLS} - \beta = \frac{\widehat{cov}(z,u)}{\widehat{cov}(z,x)} = \frac{\widehat{cov}(z,u)}{\rho\widehat{cov}(z,u) + \widehat{cov}(z,\eta)} \tag{4}$$

Note that if $\rho > 0$ then $\hat{\beta}_{2SLS} - \beta$ is the ratio of two mean zero random variables that are positively correlated. The positive correlation causes the median of this ratio to be positive, simply because the numerator and denominator are more likely than not to have the same sign. Thus, the median of $\hat{\beta}_{2SLS}$ is biased in the same direction as OLS.

Furthermore, the two random variables that determine $\hat{\beta}_{2SLS} - \beta$ in (4) both have normal distributions in large samples. Marsaglia (2006) shows such a ratio has a Cauchy distribution shifted right by $\rho Var(u)/Var(x)$, which is exactly the OLS endogeneity bias. Thus, when $\pi = 0$, we see that the median bias of 2SLS is exactly equal to the OLS bias.[4]

Among these properties, the bias of the median 2SLS estimate toward OLS has received substantial attention in the applied literature since Bound et al. (1995). The non-existence of moments and non-normal shape of the $\hat{\beta}_{2SLS}$ distribution have received substantial attention from theorists (Richardson 1968, Sawa 1969, Phillips 1983, Rothenberg 1984), and Nelson and Startz (1990) and Mikusheva (2013) provide nice expositions. But, we argue, the problems created by the association between 2SLS estimates and their standard errors have received too little attention. We will explore that issue in detail below.

As we will now explain, the four problematic finite-sample properties of 2SLS remain relevant even in large samples if instruments are weak, but they vanish if instruments are sufficiently strong. All four properties arise from the perverse influence of the sample covariance $\widehat{cov}(z,e) = \widehat{cov}(z,\eta) + \rho\widehat{cov}(z,u)$ on the denominator of (3), so it is natural to assume they will vanish in a samples large enough so that $\widehat{cov}(z,e) \approx 0$. In fact, this describes the view of applied researchers prior to Bound et al. (1995).

The error in this logic is that, in a huge sample, it is also true that z may appear to be a significant predictor of x (at the 5% level) even if the true value of π is very small. In fact, as the sample size gets larger, the value of π that is likely to render z significant at the 5% level in the first-stage of 2SLS gets small exactly as fast as $\widehat{cov}(z,e)$ gets small. As a result, if z is only significant at the 5% level (and not better), then $\widehat{cov}(z,e)$ remains non-negligible relative to $\pi\widehat{var}(z)$. Hence the perverse influence of the $\widehat{cov}(z,e)$ term on the denominator $\pi\widehat{var}(z) + \widehat{cov}(z,e)$ remains important regardless of sample size.

[4] The Cauchy has fat tails, and its mean and variance do not exist. The 2SLS estimator inherits these properties, so the distribution of $\hat{\beta}_{2SLS} - \beta$ departs seriously from normality, rendering t-tests misleading.

Thus, large samples alone are not adequate for 2SLS to have nice properties. If our instrument is significant at ***only*** the 5% level in a large sample, it is not adequate to make the problems with 2SLS vanish. Given a "weak" instrument the problematic finite-sample properties of 2SLS remain relevant even in very large samples. What we need is an instrument that is "strong," meaning it meets a higher standard of instrument relevance – see Stock et al. (2002). To explain intuitively what is meant by "strong," we start by comparing the population and sample correlations between z and x:

$$corr(z,x) = \frac{\pi Var(z)+cov(z,e)}{\sigma_z \sigma_x} = \frac{\pi Var(z)}{\sigma_z \sigma_x}, \quad \widehat{corr}(z,x) = \frac{\pi \widehat{var}(z)+\widehat{cov}(z,e)}{\hat{\sigma}_z \hat{\sigma}_x}$$

Notice how the sample correlation between z and x is driven *both* by $\pi\widehat{var}(z)$, which reflects a true relationship between z and x, and also by $\widehat{cov}(z,e)$, which reflects *spurious* correlation between z and x in finite samples. An intuitive notion of a "strong" instrument is that $\pi var(z)$ should be large enough that we are confident the sample correlation between x and z mostly reflects their true relationship, not spurious correlation that arises because $\widehat{corr}(z,e) \neq 0$ in finite samples. That is, we want $\pi var(z)$ to be large enough that we can be confident that $|\pi\widehat{var}(z)| \gg |\widehat{cov}(z,e)|$.

It is simple to see why the strange finite sample properties of 2SLS that we discussed earlier vanish if $|\pi\widehat{var}(z)| \gg |\widehat{cov}(z,e)|$. In that case, the $\pi\widehat{var}(z)$ term in the denominator of (3) dominates the $\widehat{cov}(z,e)$ term, so realizations of $\widehat{cov}(z,x)$ that are near zero become extremely unlikely. This renders non-existence of the 2SLS estimator's mean and variance a mere academic curiosity. Furthermore, if the $\widehat{cov}(z,e)$ term is negligible, (3) reduces to just $\hat{\beta}_{2SLS} - \beta \approx \widehat{cov}(z,u)/\pi\widehat{var}(z)$, which is much simpler to deal with (as it resembles the expression for OLS). Under a fixed instrument assumption, the asymptotic distribution of $\hat{\beta}_{2SLS}$ is approximately normal and centered on β. So 2SLS is "approximately" unbiased, and normality is a decent approximation to its sampling distribution.

But how can we be confident that $|\pi\widehat{var}(z)| \gg |\widehat{cov}(z,e)|$ when π and $\widehat{cov}(z,e)$ are unobserved? First, note that we can rewrite this expression as:

$$|\pi| \cdot \widehat{var}(z) \gg \hat{\sigma}_z\hat{\sigma}_e \cdot |\widehat{corr}(z,e)| \rightarrow \frac{|\pi|\hat{\sigma}_z}{\hat{\sigma}_e} \gg |\widehat{corr}(z,e)|$$

If the instrument z is valid, $corr(z,e) = 0$, so $\widehat{corr}(z,e)$ converges to zero at a $\sqrt{N}$ rate, and $|\widehat{corr}(z,e)|$ is bounded in probability by $\frac{k}{\sqrt{N}}$ for a positive constant $k{>}0$. Thus:

$$\frac{|\pi|\hat{\sigma}_z}{\hat{\sigma}_e} \gg \frac{k}{\sqrt{N}}$$

Finally, substituting our consistent first-stage estimate $\hat{\pi}$ for the unobserved π, and squaring both sides, we obtain:

$$\sqrt{N}\frac{|\hat{\pi}|\hat{\sigma}_z}{\hat{\sigma}_e} \gg k \rightarrow N\frac{\hat{\pi}^2\hat{\sigma}_z^2}{\hat{\sigma}_e^2} \gg k^2$$

Recall that the sample F statistic for significance of z in the first stage is $\hat{F} = N\hat{\pi}^2\hat{\sigma}_z^2/\hat{\sigma}_e^2$. Thus our intuitive notion of wanting confidence that $|\pi\widehat{var}(z)| \gg |\widehat{cov}(z,e)|$ corresponds to a desire to have a first-stage F-statistic that is "big" in some sense.[5] The weak in-

[5] Because $F = NR^2/(1-R^2)$, a key insight is that properties of 2SLS do not depend on N or first-stage R^2 *per se*, but only how they combine to form F. So a large sample size alone is not sufficient to ensure 2SLS has an approximately normal sampling distribution. As Mikusheva (2013) explains, the convergence of $\sqrt{N}(\hat{\beta}_{2SLS} - \beta)$ to normality as $N \rightarrow \infty$ is very slow when the first-stage R^2 is small.

strument testing literature asks just how "big" the first-stage F needs to be for 2SLS to have nice properties. We explore this literature in the next section.

3. A SIMPLE GUIDE TO WEAK INSTRUMENT TESTS

Having explained the intuition behind weak instrument tests, we now examine them in more detail. Consider a model with a single endogenous variable x and 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. We let π determine the strength of the instrument, while $\rho \in [0,1]$ controls the extent of the endogeneity problem:

$$\begin{gathered} y_i = \beta x_i + u_i \\ x_i = \pi z_i + e_i \quad \text{where} \quad e_i = \rho u_i + \sqrt{1-\rho^2}\eta_i \\ u_i \sim iidN(0,1), \eta_i \sim iidN(0,1), z_i \sim iidN(0,1) \end{gathered} \tag{5}$$

This *iid* normal setup is not as restrictive as it appears, 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.

Say we estimate the first-stage equation for x by OLS, and obtain $\hat{\pi}$ and $\hat{\sigma}_e^2$. We can test if z is a significant predictor of x using a standard F-test, given by $\hat{F} = N\hat{\pi}^2\hat{\sigma}_z^2/\hat{\sigma}_e^2$. For example, if N=1000 we conclude z is significant at the 5% level if $\hat{F} > 3.85$. This corresponds to a t-test of $t > 1.96$ (as F is the square of t in this case).

Prior to Bound et al. (1995), passing such an F-test would have been considered sufficient evidence to conclude one's instrument was relevant, and to proceed with 2SLS. But Bound et al. drew attention to situations where instruments are significant at conventional levels in the first stage of 2SLS, despite having a small or "weak" correlation with the endogenous variable. For instance, a quantitatively small correlation may be highly statistically significant in very large samples. In such cases median 2SLS estimates are biased towards OLS, and the 2SLS t-test is highly non-normal.

In an important paper, Staiger and Stock (1997) show that for an instrument to be "strong" enough for 2SLS to have acceptable finite sample properties it must meet a higher standard than 5% significance in the first stage. They formalize our statement in Section 2 that we want the first-stage F statistic to be large enough that we are confident that $|\pi\widehat{Var}(z)| \gg |\widehat{cov}(z,e)|$, which in turn implies 2SLS will have nice properties.

To understand the Staiger-Stock approach, we define the "concentration parameter" C, which measures strength of the instrument in first stage. It is closely related to the first-stage F-statistic. In particular, C is the true value of the F statistic that we could construct if we observed the unknown π and σ_e^2 that we estimate in the first stage. The finite sample properties of 2SLS only depend on N through C:

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

The sample F-statistic is an estimate of C. Just as we showed for F, if C gets large we can be confident that $|\pi\widehat{Var}(z)| \gg |\widehat{cov}(z,e)|$, and the problems with the 2SLS estimator vanish. So if C is "large" in some sense the instruments are "strong." But how large does C need to be for 2SLS to have nice properties? And how large does the sample $\hat{F}$ need to be to give confidence that the unobserved C meets that threshold?

To address these questions, Stock and Yogo (2005) focus on a particular property of 2SLS, the size of a 5%-level two-tailed t-test. This is the rate of rejecting H_0:β =0 when it is in fact true. If a test has correct size it should reject at a 5% rate. But given the non-normality of the 2SLS estimator, the t-test size may depart substantially from 5%.

Formally, Stock-Yogo derive a formula for power of the t-test in terms of C, ρ and true β that we present in Appendix A. Evaluating power at β=0 gives the size of the test. Size depends on ρ, so Stock-Yogo focus on the *maximal* size distortion, which occurs when $\rho = \pm 1$. The integral in (A3) can be evaluated numerically at different levels of C.

For example, suppose you find it acceptable that a 2SLS two-tailed t-test rejects the null H_0:β=0 at the 5% level no more than 15% of the time. In other words, you are willing to tolerate a maximal size distortion of 10%. Numerical evaluation of (A3) reveals you need C=1.82. Suppose instead you want a maximal size distortion of just 5% (i.e., your 5% t-test rejects H_0:β=0 no more than 10% of the time). Then you need C=5.78. Finally, Lee et al. (2020) show one needs C=141.6 to bring maximal size distortion down to zero.

Thus, by requiring C to be large enough, we can render the maximal size distortion as small as desired. But we can't observe C, so we must rely on sample $\hat{F}$ as an estimate. Unfortunately, because C equals $Var(z\pi)/\sigma_e^2$ times a factor of N, sample $\hat{F}$ is not a very accurate estimate of C, and it doesn't get more accurate as sample size increases.

In particular, *regardless* of sample size, the sample $\hat{F}$ is a draw from a non-central F with non-centrality parameter C. Hence, to be confident (at the 95% level) that C is at least 1.82, we need $\hat{F}$ to be at least 8.96. If we want to be confident (at 95%) that C is at least 5.78, we need to have $\hat{F}$ of at least 16.38. In general, to be confident the concentration parameter C is at least c, we need a first-stage $\hat{F}$ well above c.

Table 1 lists various levels of C, the levels of π required to achieve them if N=1000, and the first-stage F-test critical value for a 5% test that C attains the desired level. For example, to be 95% confident that C is at least 2.3, we need $\hat{F} > 10$, which corresponds to the popular Staiger-Stock rule of thumb for acceptable instrument strength.

Table 1. First-Stage F Critical Values Required to Achieve Different Objectives

Concentration Parameter ("True First-Stage F")	Value of π	F critical value to reject C<c at 5%	Rejection rate for a 5% t-test of H_0:β=0
1.82	0.0427	8.96	15%
2.30	0.0480	10.00	SS Rule of Thumb
3.84	0.0620	13.00	–
5.78	0.0760	16.38	10%
10.00	0.1000	23.10	–
73.75	0.2716	104.70	5%

Table 1 also shows, in some key cases, the (maximal) 5% two-sided t-test size achieved by that level of C. For example, if C=5.78 (which we test for using $F_{.05}$=16.38) a 2SLS 5% two-tailed t-test rejects a true null hypothesis at no more than a 10% rate.[6] The

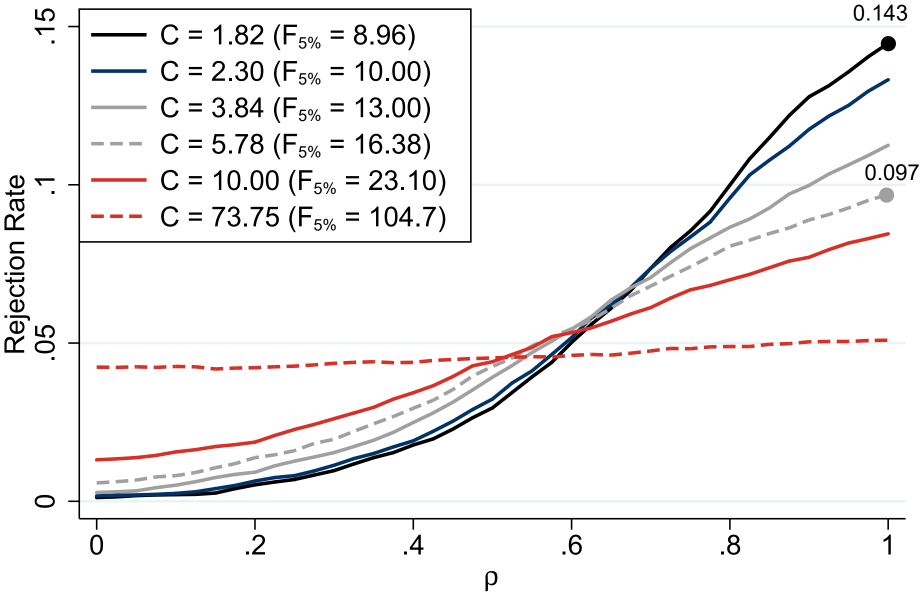
[6] We include C=10 as Staiger and Stock (1997) and Angrist and Pischke (2008) derive formulas indicating the bias of 2SLS relative to OLS is roughly $1/[1+C]$ if the first moment exists. So C=10 makes relative bias only 10%. This result is only useful with 2 or more instruments, so the mean of the 2SLS estimator exists. But we find it interesting to examine behavior of the 2SLS median in the C=10 case.

t-test size distortion depends only on C, not on N and π *per se*, so how we set N and π in Table 1 is somewhat arbitrary (i.e., the π levels are specific to N=1000). Finally, we include $F_{.05}$=104.7 in the last row of Table 1, as Lee et al. (2020) show that $\hat{F}$ >104.7 insures the maximal size of the 5% level t-test is no greater than 5%.[7]

Next, to gain a better understanding of how weak instrument tests work in practice, we implemented a simple simulation experiment to assess how well 2SLS estimates perform under each scenario in Table 1. We simulate data from the model in equation (5), assuming $\beta = 0$, varying the degree of endogeneity as captured by ρ in small increments from 0 to 1. We set the parameter π to each alternative level listed in Table 1, in order to vary the strength of the instrument. We generate artificial data sets of size N=1,000 for each combination of (π, ρ). As we've noted, it is the level of C, not N or π *per se*, that drives the properties of 2SLS. We report results from 10,000 Monte Carlo replications.

Figure 1 reports the rejection rate of a two-tailed t-test test of H$_0$: β=0 at the 5% level, based on the 2SLS parameter and standard error estimates, for each combination of C and ρ. The label shows both C and the associated F-test 5% critical level.

Figure 1. Rejection rate of $\text{H}_0 : \beta = 0$ using a 5% Two-tailed t-test



To understand Figure 1, it is important to recall that Stock and Yogo (2005) calculate worst-case (maximal) rejection rates over all values of ρ. As we see in Figure 1, the worst case corresponds to ρ near 1, so the endogeneity problem is very severe. The agreement between the results in Figure 1 and the Stock-Yogo analysis is striking. For C=1.82 they predict a worst-case rejection rate of 15%, while our simulations show 14.5%. And for C=5.78 they predict a worst-case rejection rate of 10%, while we obtain 9.7%.

But Figure 1 shows a focus on ρ=1 is not innocuous, as rejection rates vary substantially with ρ. In the next section we discuss the implications of this phenomenon.

[7] This is based on a slightly different mode of analysis from Stock-Yogo, in which the maximum of equation (A3) is now taken over C and ρ. We discuss their approach in Section 10.

4. PROBLEMS WITH 2SLS HIDDEN BY WEAK INSTRUMENT TESTS

An obvious limitation of the Stock-Yogo analysis is that the size distortion in the 2SLS t-test when ρ=1 tells us little about rejection rates at lower levels of ρ. As we see in Figure 1, the size distortion in the t-test is sharply increasing in ρ even in cases that easily pass standard tests to rule out weak instruments, such as C=10 (F_{05}=23.1). Only in the very strong instrument case of C=74 (F_{05}=105) is size roughly invariant to ρ.

The dependence of 2SLS t-test size on the nuisance parameters ρ and C reflects the fact that it is not a "pivotal statistic." It 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.

Henceforth, we will refer to instruments as "weak" if they fall in the gray area between roughly C=2.3 and C=74 where (i) they pass conventional weak IV tests, but (ii) the distribution of the t-statistic (and size of the t-test) is still highly dependent on ρ.

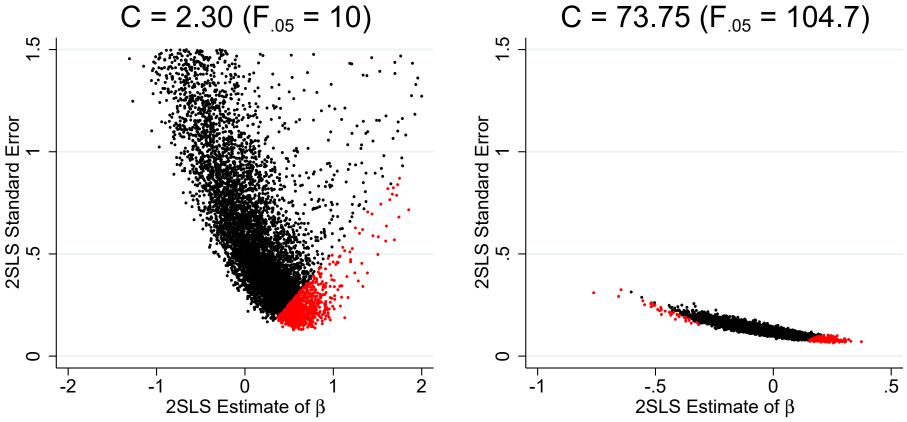
The results in Figure 1 raise serious concerns about the behavior of 2SLS t-tests when instrument are "weak" in this expanded sense. To see why, we need to understand why rejection rates are strongly increasing in ρ in such cases. There are two reasons:

First, when ρ=0 the 2SLS estimator has very low power. Estimates are roughly centered at zero, standard errors are very large, and it is very rare to reject H$_0$: β=0.

Second, if ρ>0, then, in samples where realized $\widehat{cov}(z,u)$ is high, 2SLS tends to generate both high estimates of β and (artificially) low standard errors. Thus, 2SLS obtains spurious power from the finite sample correlation between z and u. This is a very unfortunate property, as ***2SLS estimates appear to be spuriously more precise in samples where they are most shifted in the direction of the OLS endogeneity bias***. This is what causes the t-test to reject H$_0$:β=0 more frequently as ρ increases.

Figure 2 shows the association between 2SLS estimates and their estimated standard errors is quantitatively important. It plots $se(\hat{\beta}_{2SLS})$ against $\hat{\beta}_{2SLS}$. The left panel shows the case of ρ=0.8 and C=2.30 ($F_{.05}$=10). A strong negative association is very evident; in fact, the Spearman r_s is -0.576 and Kendall's τ is -0.511. The magnitudes involved are also substantial – 2SLS estimates that are most shifted toward the OLS bias appear to be *much* more precisely estimated. Of course this appearance is spurious.

Figure 2. 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.*

The red dots in Figure 2 indicate cases where $\hat{\beta}_{2SLS}$ differs significantly from zero according to a two-tailed 5% t-test. In the C=2.30 ($F_{.05}$=10) case, the hypothesis H$_0$: $\beta = 0$ is rejected at a 10% rate. Due to the negative association between the 2SLS estimates and their standard errors, all rejections occur when $\hat{\beta}_{2SLS}$>0, and none when $\hat{\beta}_{2SLS}$<0. Only the estimates most shifted towards the OLS bias are ever judged significant.

The right panel of Figure 2 shows the case of ρ=0.8 and C=74 ($F_{.05}$=105). When the instrument is this strong the negative association between $se(\hat{\beta}_{2SLS})$ and $\hat{\beta}_{2SLS}$ persists. In fact, Spearman's r_s is -0.92 and Kendall's τ is -0.75, showing this is not just a weak instrument phenomenon. 2SLS now has a rejection rate near the correct 5% rate (4.87%). But 93% of those rejections occur when $\hat{\beta}_{2SLS} > 0$. A one-tailed 2.5% test of H$_0$: $\beta \leq 0$ rejects at a 4.54% rate. The asymmetry in positive vs. negative rejections is a direct consequence of the negative association between 2SLS estimates and standard errors.

Table 2 gives a broader view of this phenomenon by showing Spearman's r_s between $se(\hat{\beta}_{2SLS})$ and $\hat{\beta}_{2SLS}$ for different levels of C and ρ. Clearly, the negative association is not specific to the examples in Figure 2. Table 2 shows how this relationship gets stronger as ρ increases. This drives the pattern of rejection rates increasing with ρ in Figure 1.

Table 2. Spearman Rank Correlations (r_s) between $se(\beta_{2SLS})$ and $\hat{\beta}_{2SLS}$

Concentration	$F_{5\%}$	Spearman Correlations (r_s)		
Parameter		$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
1.82	8.9	-0.113	-0.291	-0.492
2.3	10	-0.133	-0.359	-0.576
5.78	16.38	-0.267	-0.612	-0.875
73.75	104.7	-0.350	-0.720	-0.917

We now study power of the 2SLS t-test in more detail by simulating probabilities of rejecting H$_0$:β=0 when it is false. We consider alternative versions of model (5) where the true β is 0.30 or -0.30. Importantly, these would be quantitatively large but plausible values in typical empirical applications, as they imply a one standard deviation change in x induces an 0.25 standard deviation change in y. The results are reported in Table 3.

Table 3. Power of 2SLS t-Test – Frequency of Rejecting H$_0$: $\beta = 0$ (%)

Concentration	$F_{5\%}$	$\beta = 0.3$			$\beta = -0.3$		
Parameter		$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$
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
3.84	13.00	4.4	15.9	25.1	4.2	0.3	1.7
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: $\beta = 0$.

A striking result in Table 3 is that the 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 thresholds. For example, in the C=2.3 ($F_{.05}$=10) case widely considered an acceptable threshold for a strong instrument, the probability of rejecting the false null H_0: β=0 is only 0.2% when the true β is -0.3 and $\rho = 0.50$. Increasing instrument strength substantially to C=10 ($F_{.05}$=23.1) only increases power to 2.3%.

The negative association between 2SLS estimates and standard errors that arises when the OLS bias is positive drives this result. This is evident from the geometry of Figure 2. If β=-0.30 the cloud of points in Figure 2 is shifted left. This shift moves the most precisely estimated $\hat{\beta}$'s closer to zero, so they are rarely significant.

Another clear pattern in Table 3 is that power of the 2SLS t-test is asymmetric when the OLS bias is positive ($\rho > 0$). For example, if C=10 and ρ=0.5 the probability of rejecting the false null H_0: β=0 is 23.7% when $\beta = 0.30$ compared to only 2.3% when β=-0.30. This asymmetry also arises from the geometry of of Figure 2. If β=0.30 the cloud of points in Figure 2 is shifted right. This clearly generates more significant results, in contrast to the leftward shift (β=-0.30) that generates fewer.

Appendix A presents an expanded analysis of 2SLS t-test power curves. This shows the power asymmetry we have described is quite a general phenomenon. 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.

Finally, focusing on all the "weak" instrument cases in Table 3 – by which we mean all cases except C=74 – we see that the power of the 2SLS t-test is very low when ρ=0. But if true β is positive then power increases with the degree of endogeneity ρ. This is because, as ρ increases, the t-test derives more spurious power from the finite sample correlation between the instrument z and the structural error u.[8]

4.1. Median Bias of 2SLS

A potential alternative explanation for the pattern in Figure 1 is that bias in the median 2SLS estimate increases as the degree of endogeneity (ρ) increases. In principle, this could cause the rate of rejecting H_0: $\beta = 0$ to increase with ρ. But we can rule this out. For all values of C we consider, the instruments are strong enough that median bias in 2SLS is negligible, or at least modest, regardless of the degree of endogeneity.

Figure C4 illustrates this point. If C=10 or better, 2SLS estimates are essentially median unbiased. Figure A4 plots the median 2SLS bias *relative* to the OLS bias. Relative bias is modest in all cases. For example, when C=2.30 ($F_{.05}$=10) the 2SLS median bias is less than 15% of the OLS bias unless ρ is very small. Thus, if median bias were one's only concern, a first-stage $\hat{F}$ of 10 would be quite sufficient.

4.2. Summary

The Stock-Yogo test assesses worst-case size distortions in two-tailed 2SLS t-tests. If an instrument passes these tests, it implies that size distortions are modest. But this conceals serious power problems that afflict the t-test unless instrument strength is far above conventional thresholds. *It is an unfortunate property of the 2SLS estimator 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.* One important consequence is the t-test has little power to detect plausibly large true negative effects when the OLS bias is positive. 2SLS is approximately median unbiased, but this property is not very useful if only estimates shifted in the direction of the OLS bias are likely to be significant.

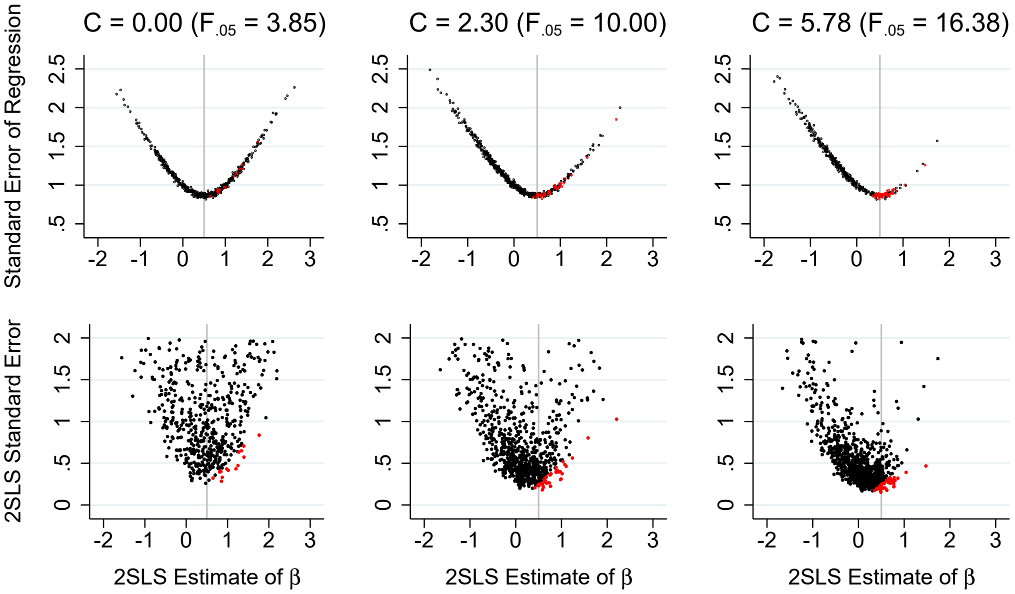
[8] Positive realizations of $\rho\widehat{cov}(z,u)$ increase $\widehat{cov}(z,x)$, the sample covariance between the instrument and the endogenous variable. This, in turn, generates a spurious reduction in the 2SLS standard error.

5. INSIGHTS FROM FINITE SAMPLE THEORY

Results from finite sample theory can help us understand the strong dependence bewteen 2SLS estimates and their standard errors. Phillips (1989) derives two key properties of 2SLS in the unidentified case where the instrument is irrelevant. First, the 2SLS estimator converges in distribution to a scale mixture of normals centered on $E(\hat{\beta}_{OLS})$. Second, the 2SLS variance estimator ($\hat{\sigma}^2$) converges in distribution to a quadratic function of $\hat{\beta}_{2SLS}$, with a minimum at $E(\hat{\beta}_{OLS})=\rho Var(u)/Var(x)$.

These properties are shown in Figure 5, obtained by applying 2SLS to 1000 datasets of size N=1000 from the DGP in eqn. (5). We set β=0 and ρ=0.50 so $E(\hat{\beta}_{OLS})$=0.50. The left panel shows results in a completely unidentified model (C=0,π=0). As we see in the upper left, the standard error of regression ($\hat{\sigma}$) is indeed minimized at $\hat{\beta}_{2SLS}$=0.5. Of course, $\hat{\sigma}$ is a key driver of the standard error of $\hat{\beta}_{2SLS}$. This causes $\hat{\beta}_{2SLS}$ to appear (spuriously) more precise in the vicinity of the OLS bias, as we see in the lower left panel that plots $se(\beta_{2SLS})$ against $\hat{\beta}_{2SLS}$. Note how this looks similar to Figure 2.

Figure 3. The Dependence Between 2SLS Estimates and Standard Errors



Note: Runs with standard error > 2 not shown. Red dots indicate H_0 : $\beta = 0$ rejected at 5% level.

A point that is little appreciated (at least by applied researchers) is that this property of 2SLS in the unidentified case has a major influence on the behavior of 2SLS estimates and standard errors in strongly identified models. For this reason, Phillips (1989) calls this the "leading case." To illustrate, the middle and right panels of Figure report results for identified models where C= 2.3 or 5.78. As the strength of identification increases, the 2SLS estimates shift left; they move away from the OLS bias and start to become median unbiased. Strikingly however, the quadratic relationship between $\hat{\sigma}^2$ and $\hat{\beta}_{2SLS}$ is unaffected. Stronger identification merely shifts most of the 2SLS estimates into the left side of the quadratic curve, but it does not affect the curve itself.

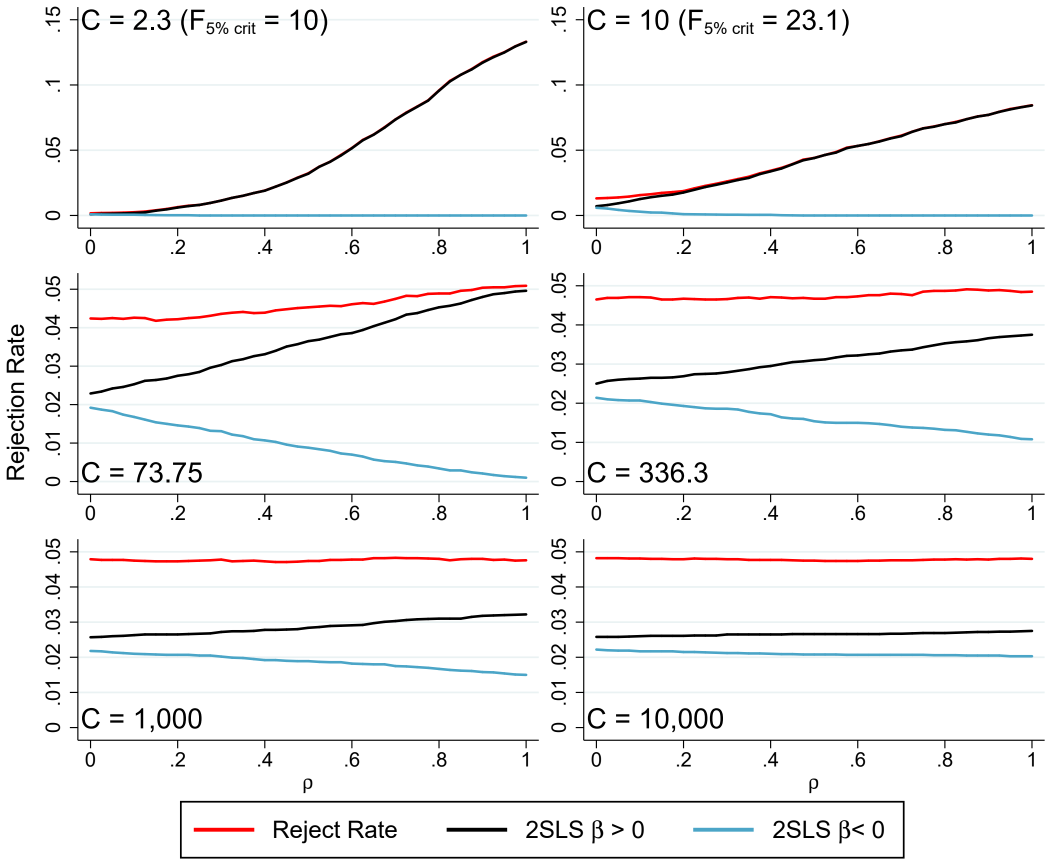
Thus, even in strongly identified models, there exists a strong association between the 2SLS standard error of regression ($\hat{\sigma}$) and the 2SLS estimate, such that ($\hat{\sigma}$) is minimized when estimates are in the vicinity of the OLS bias. This causes 2SLS estimates that are shifted towards the OLS bias to appear spuriously precise.

6. ONE-TAILED T-TESTS

An important consequence of the association between 2SLS estimates and their standard errors is that size distortions in one-tailed tests are much greater than distortions in two-tailed tests. This is shown in Figure 4. The red lines show rejection rates of two-tailed 5% t-tests of $H_0 : \beta = 0$ for different levels of C and ρ. The black/blue lines show how frequently these rejections occur at positive/negative values of $\hat{\beta}_{2SLS}$. This is equivalent to plotting rejection rates of one-tailed 2.5% t-tests of H_0: $\beta \leq 0$ and H_0: $\beta \geq 0$.

In the case of C=2.30 ($F_{.05}$=10) that corresponds to the Staiger-Stock rule of thumb for acceptably strong instruments, the rejection rate of **both** the 5% two-tailed test and the 2.5% one-tailed test against H_0: $\beta \leq 0$ increase from 0% to to 14.5% as ρ increases from 0 to 1. But the one-tailed test against the null of H_0: $\beta \geq 0$ **never** rejects.

Figure 4. Rejection Rates of One and Two-Tailed Tests



We have seen that in the strong instrument case of C=74 ($F_{.05}$=105) there is negligible size distortion in two-tailed t-tests. Indeed, in Figure 4 we see the rejection rate of a 5% two-tailed test increases only modestly from 4.1% if $\rho = 0$ to 5% when $\rho = 1$.

But the size distortions in one-tailed t-tests remain substantial: The rejection rate of a 2.5% one-tailed test against H$_0$: $\beta \leq 0$ *increases* from 2.2% to 5% as ρ increases from 0 to 1. Conversely, the rejection rate of a 2.5% one-tailed test against H$_0$: $\beta \geq 0$ *declines* from 1.9% when $\rho = 0$ down to essentially zero when $\rho = 1$. This asymmetry is a direct consequence of the negative association between $se(\hat{\beta}_{2SLS})$ and $\hat{\beta}_{2SLS}$, which imparts positive (negative) 2SLS estimates of β with spuriously high (low) precision.

How strong do instruments need to be for one-tailed t-tests to have correct size? As we see in Figure 4, even if the concentration parameter C is increased to 336.3, which corresponds to a first-stage $F_{.05}$=400, the asymmetry in rejection rates for one-tailed tests remains substantial. Strikingly, we find C must be increased to roughly 10,000 to eliminate size distortions in one-tailed 2SLS t-tests. Only then are rejection rates of one-tailed tests insensitive to the level of ρ.

The asymmetry in rejection rates of one-tailed t-tests is of great practical importance. Applied researchers almost always use two-tailed tests because symmetry makes one-tailed tests redundant (e.g., a 5% two tailed-test is equivalent to a 2.5% one-tailed test). But as we see, this is false for 2SLS, even with very strong instruments.

7. THE ANDERSON-RUBIN TEST

A common suggestion in the weak IV literature is to avoid the t-test, and instead use a test that is "robust" to weak instruments – in the sense that it has correct size even if instruments are weak. In the one endogenous variable, single instrument case the unambiguous suggestion is the Anderson and Rubin (1949) test. The AR test is simply the F-test from the reduced-form regression of y on z, which is $y = z\beta\pi+(\beta e+u) = z\xi+v$ where $\xi = \beta\pi$. Thus, the AR test judges $\hat{\beta}_{2SLS}$ to be significant if the F-test indicates that z is a significant predictor of y in the reduced form regression. Why does this work? 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 $H_0 : \beta = 0$.

The AR test has correct size, regardless of instrument strength, because it is simply an F-test from an OLS regression (i.e., it is a pivotal statistic). But AR it has other highly desirable properties as well. The superior power properties of the AR test relative to the t-test are illustrated in Figure 5. It presents analytical power curves for both tests, for the model in (5), obtained as described in Appendix A. We set the level of instrument strength to C=10, which is well above conventional weak instrument thresholds (i.e., to have 95% confidence that C is at least 10 one needs a first-stage $\hat{F} > 23.1$). 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.

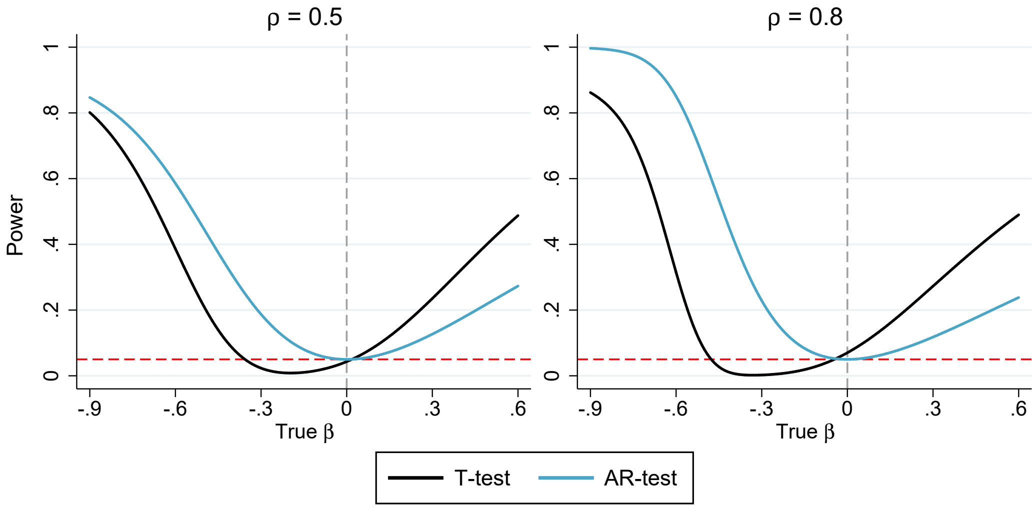
An unbiased statistical test has the desirable property that the probability of rejecting $H_0 : \beta = 0$ is minimized if the true β is in fact zero. We can see in Figure 5 that the AR test is unbiased. It also has correct size, as its power evaluated at $\beta = 0$ is exactly 5%.

In contrast, the t-test is biased. As we see in left panel of Figure 5, 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 the true β is near -0.25 rather than at zero. And if $\rho = 0.80$ (right panel) the power of the t-test is near zero for true β in the -0.25 to -0.40 range. Recall that for the model in (5), β is approximately equal to the change in y (in standard deviations) induced by a one standard deviation change x. Effect sizes in this -0.25 to -0.40 std. dev. range are quite large in typical applications. Thus, these results show that the t-test has essentially no power to detect a wide range of substantively large true negative effects when the endogeneity bias afflicting OLS is positive.

This pattern is a direct consequence of the positive association between 2SLS estimates and standard errors that we discussed in Sections 4-6, which can cause the t-test to have very low power when the true value is opposite in sign to the OLS bias. In contrast, as we can also see in Figure 5, the AR test has far better power to detect true negative effects. Crucially, Moreira (2009) shows that, in the single endogenous variable, single instrument case, the AR test is the uniformly most powerful unbiased test. This means it has better power than any other unbiased test, regardless of the true parameter value.

Figure 5 also shows that when the true β is positive the t-test has higher power than the AR test against H_0:β=0. But this property is not desirable: It reflects the facts that (i) the t-test is biased, and (ii) the 2SLS standard errors are spuriously small for estimates that are shifted in the direction of the OLS bias (positive).

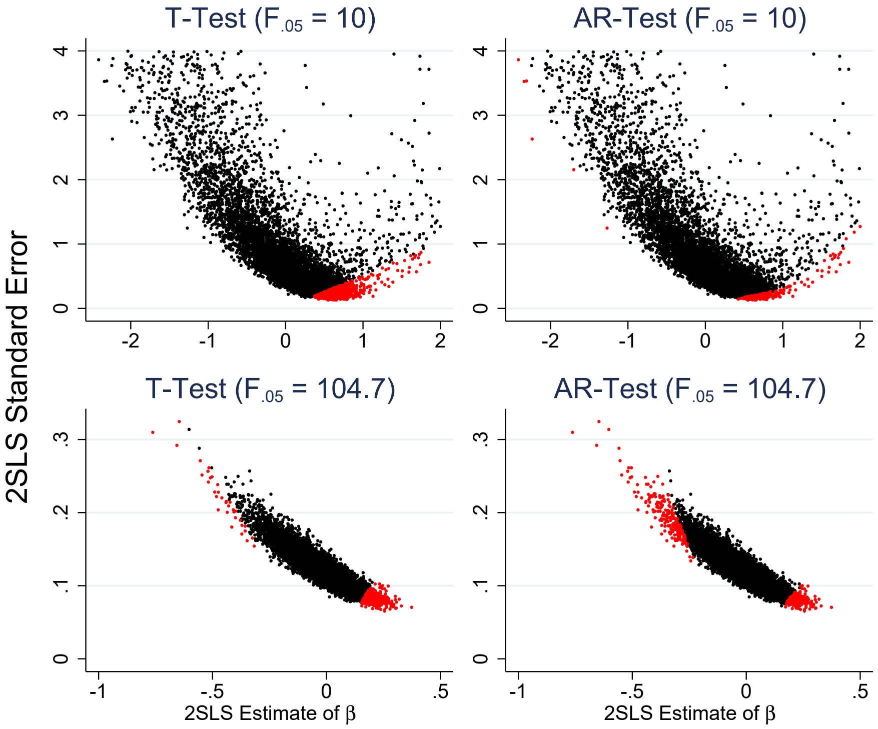
Figure 5. Power of the T-Test vs. AR-Test when $C = 10$



Another notable aspect of Figure 5 is the fact that the power of the two-tailed 5%-level t-test evaluated at $\beta = 0$ is roughly 5% when $\rho = 0.5$, and about 6% when $\rho = 0.8$. This illustrates the point of Angrist and Kolesár (2021) that two-tailed t-test size inflation is minor except in cases where endogeneity is extremely strong and/or the instrument is very weak. If this were one's only concern one might argue – as they do – for a sanguine view of the performance of the t-test. We argue instead that the truly concerning problems with the t-test are its bias and poor power properties, which the AR test avoids.

One thing Figure 5 does not reveal is the frequency of positive vs. negative rejections, as the power of a two-tailed test is defined as their sum. In Sections 4 to 6 we argued it is important to consider the sign of rejections. We consider this in Figure 6, which compares results from t-tests vs. AR tests, focusing on the case of ρ=0.80. The top panel considers the case of C=2.3 ($F_{.05}$=10) which corresponds to the Staiger-Stock rule of thumb for acceptably strong instruments. Here the 5% two-tailed t-test rejects H$_0$:β=0 at a 10% rate, and these all occur when $\hat{\beta}_{2SLS} > 0$, reflecting the severe power asymmetry of the t-test. The AR test rejects H$_0$:β=0 at a 4.8% rate, which only differs from the correct 5% rate due to sampling variation. However, in the top right of Figure 6, we see that 85% of those rejections occur when $\hat{\beta}_{2SLS} > 0$. So using AR does not avoid the asymmetry that most rejections occur at positive values, which is the direction of the OLS bias.

Figure 6. T-test vs. AR test rejections: $SE(\hat{\beta}_{2SLS})$ plotted against $\hat{\beta}_{2SLS}$ itself ($\rho = 0.80$)



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).

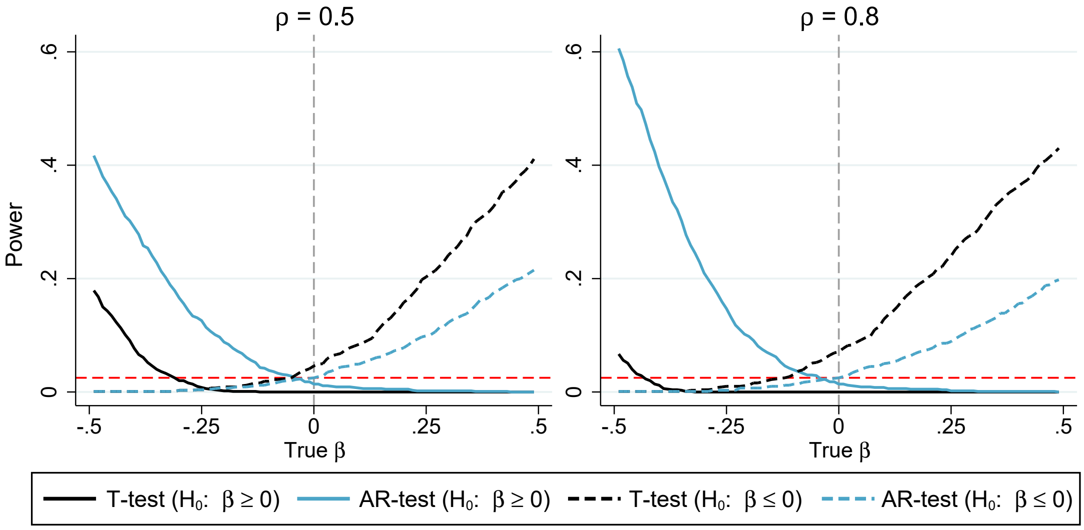
Thus, if instruments are weak, the AR test shares with the t-test the problem that it is more likely to call 2SLS estimates significant if they are shifted in the direction of the OLS bias. The reason is again the strong positive association between $\rho\widehat{cov}(z,u)$ and $\hat{\beta}_{2SLS}$. A large value of $\rho\widehat{cov}(z,u)$ also generates a large value of the AR test. Thus, if $\rho > 0$ the AR test and $\hat{\beta}_{2SLS}$ have a positive association. Hence, the AR test is more likely to reject H$_0$: $\beta = 0$ if $\hat{\beta}_{2SLS} > 0$. Importantly, however, in contrast to the t-test, this problem with the AR test vanishes quickly as instrument increase - as we now show.

The bottom panel of Figure 6 reports results for the strong instrument case of $F_{.05}$=105. Here the size distortion in the two-tailed t-test is mostly eliminated. But, as the red shading shows, 93% of those rejections occur when $\hat{\beta}_{2SLS} > 0$. In contrast, the AR test exhibits a fairly even balance of positive (54%) vs. negative rejections. Thus, the AR test achieves this balance at a vastly smaller first-stage F than required for the t-test. This provides an additional reason to prefer the AR test to the t-test.

The AR test can be cast as a one-tailed test as follows: Reject H$_0$: $\beta \leq 0$ if the AR test is significant and $\hat{\beta}_{2SLS} > 0$. Conversely, reject H$_0$: $\beta \geq 0$ if the AR test is significant and $\hat{\beta}_{2SLS} < 0$. Figure 7 compares the power of one-sided 2.5% level AR and t-tests. We look at the case of C=10 and $\rho = 0.50$ or 0.80. If the size of these four tests were correct, then each should have a 2.5% rejection rate when the true $\beta = 0$.

As we have already seen, the size distortion in one-tailed t-tests is substantial. Consider the ρ=0.80 case: When $\beta = 0$ the 2.5% one-tailed test against H$_0$: $\beta \geq 0$ has power of 12% while that against H$_0$: $\beta \leq 0$ has zero power. In fact, the power of a one-tailed t-test to detect a true negative effect is essentially zero unless the effect size is at least -0.5. In contrast, the power of the one-tailed AR tests at $\beta = 0$ are much closer to 2.5% (i.e., 3% on the positive side and 2% on the negative side). The AR test has much greater power to detect true negative effects. For example, for a true $\beta = -0.4$ the one-tailed AR test against H$_0$: $\beta \geq 0$ has power of roughly 50% compared to near zero for the t-test.

Figure 7. Power of the One-Tailed t-Test vs. AR-Test when $C = 10$



The flip side of this pattern is that the one-tailed t-test has spuriously inflated power to detect positive effects in this context (where the OLS bias is positive). The 2.5% level t-test has power of 12% to reject H$_0$: $\beta \leq 0$ when the true β=0, while the size of the AR test is almost exactly 2.5%. As we move to higher positive values of the true β, the t-test power remains above that of the AR test. We emphasize that this is not a desirable property, as it derives from the fact that the 2SLS standard errors are artificially small when the 2SLS estimate is shifted in the direction of the OLS bias, exaggerating the precision of positive estimates.

Summarizing the results of Sections 6 and 7, we have seen that the t-test causes 2SLS estimates to appear spuriously more precise when they are shifted in the direction of the OLS bias – even if instruments are strong – but the AR test is far less sensitive to this problem. Thus we make the following general recommendation: ***Based on its superior size and power properties, and to avoid over-rejecting the null when $\hat{\beta}_{2SLS}$ is shifted in the direction of the OLS bias, one should rely on the AR test rather than the t-test even when the first-stage F-statistic is in the thousands***.

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 8 and in our companion paper Keane and Neal (2021).

Second, the AR statistic is "pivotal," meaning its distribution under H$_0$: β=0 does not depend on the unknown ρ and π that govern the degree of endogeneity and strength of the instrument. This allows one to invert the AR test to form valid confidence intervals, as discussed in Anderson and Rubin (1949) and Dufour (2004), and illustrated Section 8. In contrast, the distribution of the t-statistic is highly dependent on ρ and π, rendering confidence intervals suspect even in large finite samples with weak instruments. If instruments are very weak the AR confidence interval can be unbounded. But as Dufour (2004) notes, this is not so much a problem as an accurate reflection of uncertainty.

Third, with multiple instruments the AR test is no longer optimal. Instead, Moreira (2003) shows the conditional likelihood ratio (CLR) test is the uniformly most powerful unbiased test (the AR and CLR tests are equivalent in the single instrument case). Finlay and Magnusson (2009) provide a Stata command to calculate a heterskedastcity robust version of the CLR test, and to invert it to form confidence intervals. We compare the performance of t, AR and CLR tests in the over-identified case in Section 13.

8. EMPIRICAL EXAMPLE: THE "EXCESS" SENSITIVITY OF CONSUMPTION TO CURRENT INCOME

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. As we will see, this application is characterized by a concentration parameter C that is just above 10. Thus conventional weak instrument testing thresholds are met, but as we will see, issues related to weak instruments are still relevant. It is interesting to examine the behavior of 2SLS hypothesis tests in this context.

As background, we note that simple versions of the permanent income hypothesis (PIH) imply this should be zero. A positive value is often referred to as "excess sensitivity," which may be evidence of liquidity constraints. Note, however, that elaborations of the PIH to account for consumption/leisure substitution and/or consumer prudence (i.e., reluctance to borrow against uncertain future income) may also help to explain "excess sensitivity." Regardless, the elasticity of consumption with respect to anticipated income changes is of considerable interest.

To estimate the elasticity we run the regression:

$$\Delta lnC_{it} = \alpha + \beta \Delta lnY_{it} + \boldsymbol{\gamma} \boldsymbol{V}_{it} + \epsilon_{it} \tag{6}$$

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 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 has followed a sample of over 5,000 U.S. households and their descendants since 1968. The PSID became biannual in 1999. We use the most comprehensive consumption measure,[9] available from 2005 to 2019, giving us eight observations per household. The consumption and income variables refer to the survey year, so in estimating (6) 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.[10] This process left us with 643 households and seven observations per household, for a total sample size of $4,501$.

We report the results in Table 4. Estimating (6) 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.

We report the first stage 2SLS results in the second column of Table 4. As expected $lnY_{i,t-2}$ is a highly significant predictor of ΔlnY_{it}. A higher level of 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 4. 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 we obtain a 95% confidence interval of (0.03, 1.57) 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.

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

[10] We restrict the sample to households whose income was between $\$3,000$ and $\$1,000,000$ in every year. We also dropped 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 in an attempt to reduce noise.

Table 4. 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]	
$\Delta ln Y_{t-2}$		-0.0321 (0.0100) [0.0078]		-0.0177 (0.0085) [0.0062]
F-Stat (Hetero-σ Robust)		10.283		4.312
p-value		0.0014		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 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 4 on all 10,000 datasets, and summarize the results in Table 5.

Our method of constructing samples means the point estimates in Table 4 are the true values of the data generating process in our simulation experiment, and the concentration parameter C of the DGP is 10.28. Thus, we are well above conventional thresholds for an acceptably strong instrument. We begin by noting two features of Table 5:

First, the median OLS, 2SLS and reduced form estimates all agree closely with the point estimates reported in Table 4. The mean OLS and reduced form estimates also agree, while of course the mean of the 2SLS estimates does not (as the mean of the 2SLS estimator does not exist in the exactly identified case).

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 4. Thus, the asymptotic standard errors are a good guide to the actual sampling variation of the OLS and reduced from estimates.[11] In contrast, the empirical standard deviation of the 2SLS estimates bears no resemblance to the 2SLS standard error, which is expected as the variance of 2SLS does not exist in the exactly identified case.

[11] Table 4 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.

Table 5. 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

Now we evaluate the power and size of 2SLS t-tests. In Figure 8 we plot $se(\hat{\beta}_{2SLS})$ against $\hat{\beta}_{2SLS}$ across the 10,000 samples. The association between 2SLS estimates and their standard errors is very evident. In this DGP the correlation ρ between the errors in the structural and reduced form equations is -0.40.[12] This is why the mean OLS estimate of 0.14 is less than the true elasticity of $\beta = 0.55$. As we see in Figure 8, the 2SLS standard errors imply that 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 $\beta = 0.55$.

In the top left panel of Figure 8 we shade in red cases where $\hat{\beta}_{2SLS}$ is significant at the 5% level, which occurs 39.7% of the time. In the right panel we consider two-tailed 2SLS t-tests of the null hypothesis that $\beta = 0.55$. The red dots indicate rejections at the 5% level. This occurs in 3.58% of cases, so the size of the test is too small. This is what we expect based on the case of C=10 ($F_{.05}$=23) and $\rho = 0.4$ in Figure 1. More importantly, almost all rejections occur when $\hat{\beta}_{2SLS}$ is near zero, because the 2SLS standard errors are (spuriously) smaller when the estimate is shifted in the direction of the OLS bias.

The bottom panel of Figure 8 reports the same results except now we use the AR test to evaluate significance of the 2SLS estimates. 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. In the right panel we consider AR tests of the null hypothesis that $\beta = 0.55$. This is simply the 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.[13]

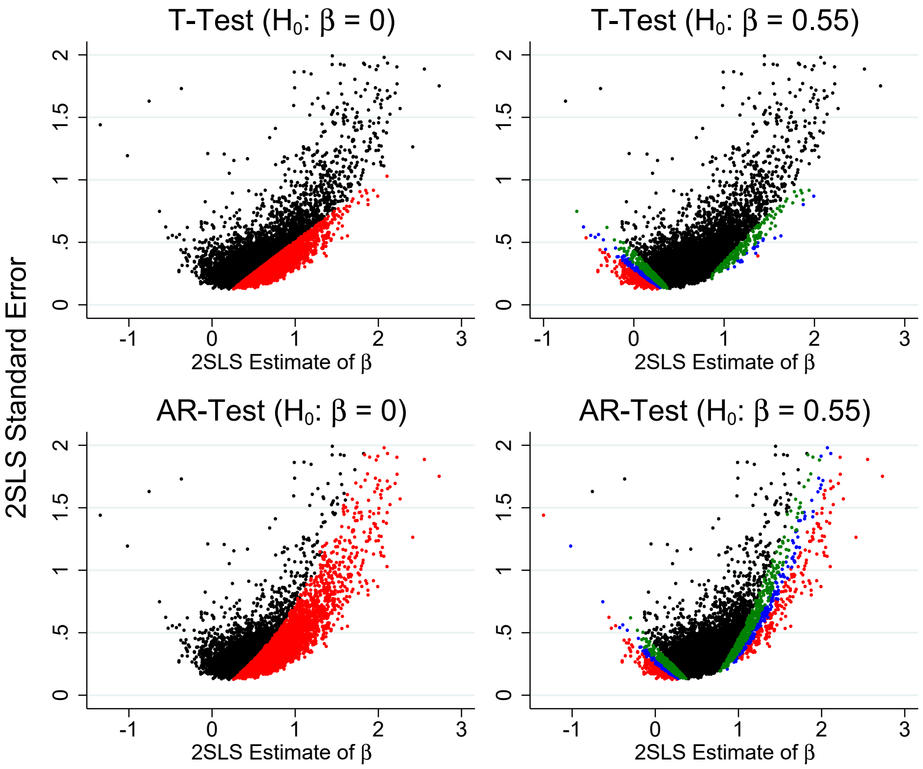
These results show that the AR test exhibits both substantially better power and more accurate size than the t-test in this data environment. Moreover, it does not suffer from the problem that estimates shifted in the direction of the OLS bias appear to be more precise. This illustrates that the problems with 2SLS t-tests and advantages of the AR test that we illustrated in Sections 4–6 are not limited to the *iid* normal environment but are also evident in a realistic environment constructed from actual data. In Keane and Neal (2021) we present another application to estimating labor supply elasticities. In that case $\rho = 0.70$, and the advantages of the AR test are much greater.

The AR test is widely recommended by econometric theorists as superior to the t-test when instruments are weak, and no worse when instruments are strong. For example, Andrews et al. (2019) argue the AR test should be widely adopted by applied researchers,

[12] A negative ρ generates a positive association between 2SLS estimates and standard errors.

[13] 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.

Figure 8. Standard Error of $\hat{\beta}_{2SLS}$ plotted against $\hat{\beta}_{2SLS}$ itself



Note: Runs with standard error > 1 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.

stating: "In just-identified models ... Moreira (2009) shows that the AR test is uniformly most powerful unbiased.... Thus, the AR test has (weakly) higher power than any other size-α unbiased test... In the strongly identified case, the AR test is asymptotically efficient... and so does not sacrifice power relative to the conventional t-test. ... Since AR confidence sets are robust to weak identification and are efficient in the just-identified case, there is a strong case for using these procedures in just-identified settings."

Despite its clear advantages, the AR test has been largely neglected by applied researchers. In fact, as far as we know, it has never been adopted in the large literature on estimating the elasticity of consumption with respect to anticipated income changes. In our application, given that the first-stage F statistic is only slightly above 10, conventional wisdom says we are in a borderline case where weak instruments may or may not be a concern. Clearly the AR test should be viewed as more reliable than the t-test in this context. Our experiment illustrates just how superior the AR test is in practice.

A limitation of the AR test is that it loses power in over-identified settings, where Moreira (2003)'s conditional likelihood ratio (CLR) test may be preferable. We compare them in Section 13. But the the AR and CLR tests are equivalent in the just-identified case considered here. In the next section we discuss the ideas behind "conditional" tests, and focus on the conditional t-test, which can be useful in the just-identified case.

9. THE CONDITIONAL T-TEST APPROACH

As we have seen, 2SLS t-tests can give misleading results due to the association between 2SLS estimates and standard errors, even when instruments are strong by conventional standards. This same basic problem affects the AR test, but much less severely, making it clearly preferable to the t-test in the single instrument case. Moreira (2003) and Mills et al. (2014) have proposed a different approach to 2SLS hypothesis testing known as the "conditional t-test" approach. It provides another useful way to address this problem.

Due to the association between 2SLS estimates and standard errors, the distribution of 2SLS t-tests under the null is not well approximated by a $N(0,1)$. In fact, it is highly asymmetric, even when instruments are strong. The idea of a conditional t-test is to adjust the critical values of the 2SLS t-test to take this non-normality and asymmetry into account. This is achieved by using critical values that are conditional on the first-stage F and the covariance matrix of the errors in the reduced form regressions of y and x on z. These are $y = z\beta\pi + (\beta e + u)$ and $x = z\pi + e$. Under $\mathrm{H}_0 : \beta = 0$ we have $y = u$ so the covariance of the reduced form errors provides an estimate of $\rho = cov(e, u)$.

As we saw in Figure 1, the size of 2SLS t-tests using standard critical values depends on F and ρ. Not surprisingly then, it is possible to invert this relationship to find appropriate critical values conditional on F and ρ such that the t-test has the correct size. As Mills et al. (2014) note however, there is no known closed form solution for this inversion, so it must be calculated by simulating the distribution of the 2SLS t-test conditional on estimates of F and ρ. Fortunately this is a simple process – see B.[14]

Table 6 reports summary statistics for critical values simulated using the DGP in (5) assuming a true $\rho = 0.80$. For each level of $C(F_{.05})$ we report the median and standard deviation of the calculated critical values. Given each dataset drawn from our DGP, we get estimates of F and ρ, and then simulate the distribution of the t-statistic conditional on $(\hat{F}, \hat{\rho})$. Thus the construction of Table 6 requires running a simulation within a simulation. Hence, below each median critical value, we report in parenthesis the standard deviation of the critical values constructed under each true $C(F_{.05})$ scenario.

For example, in the case of C=2.30 ($F_{.05}$=10) that corresponds to the Staiger-Stock rule of thumb, the median critical values for 2.5% left-tailed and right tailed t-tests are -0.445 and 3.152, respectively. Using these critical values, one-tailed tests have the correct 2.5% size. The large deviation of these values from ±1.96 shows the extreme asymmetry generated by the negative association bewteen 2SLS estimates and standard errors in this case. As Mills et al. (2014) note, the left and right tail conditional critical values can be used in conjunction to form a two-tailed 5% conditional t-test with correct size.

Figure 9 presents a graphical illustration of how the asymmetric two-tailed conditional t-test (henceforth "ACT") works. The left panel shows results for 10,000 simulated datasets with C=2.30 ($F_{.05}$=10) and $\rho = 0.80$. As before, we plot $SE(\hat{\beta}_{2SLS})$ against $\hat{\beta}_{2SLS}$. The red dots indicate cases where we reject the null H_0:$\beta = 0$ because the ratio of $\hat{\beta}_{2SLS}$ to the conditional critical value exceeds one. The ACT achieves a correct overall 5% rejection rate, as well as symmetry with 2.5% negative and 2.5% positive rejections.

We have observed that standard 2SLS t-tests have little power to detect true negative effects when the OLS bias is positive, even when instruments are "strong" by conventional

[14] Basically, one draws many datasets with the same $(\hat{F}, \hat{\rho})$ but with different structural errors, using a construction in Moreira (2003). One then simulates the distribution of the t-test across these artificial samples as we explain in Appendix B. Marcelo Moreira provided us with his Matlab code which does this calculation efficiently.

Table 6. Median Critical Values for the Mills et al. (2014) One-Tailed t-tests

	1%	2.5%	5%	95%	97.5%	99%
$C = 2.3$	-0.446	-0.445	-0.441	2.606	3.152	3.745
	(0.176)	(0.176)	(0.175)	(0.132)	(0.126)	(0.118)
$C = 10$	-0.928	-0.921	-0.899	2.253	2.760	3.338
	(0.173)	(0.161)	(0.141)	(0.093)	(0.114)	(0.133)
$C = 73.75$	-1.774	-1.561	-1.377	1.908	2.306	2.760
	(0.039)	(0.027)	(0.020)	(0.017)	(0.023)	(0.031)
$C = 336.3$	-2.048	-1.754	-1.516	1.781	2.132	2.526
	(0.009)	(0.007)	(0.005)	(0.005)	(0.006)	(0.009)
$C = 1,000$	-2.147	-1.823	-1.567	1.732	2.064	2.434
	(0.004)	(0.003)	(0.002)	(0.002)	(0.003)	(0.004)
$C = 10,000$	-2.239	-1.888	-1.613	1.685	1.998	2.344
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

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

standards (e.g., F=10). This is of great practical importance, as it means there is little chance of detecting negative program effects given positive selection on unobservables. We see here that the ACT solves this problem by using a very "lenient" critical value in the left tail (in this case approximately -0.445 instead of -1.96).[15] Applied researchers may find it odd to adopt such a "weak" standard, but, given that association between 2SLS estimates and standard errors is an intrinsic property of the estimator, it is essential if one desires a "first do no harm" approach to policy evaluation. Conversely, the ACT adopts a very "strict" critical value of 3.152 for assessing positive effects.

The right panel of Figure 9 presents results for the relatively strong instrument case of C=74 ($F_{.05}$=105). Here the median left and right tail critical values used to form the 5% two-tailed ACT are -1.561 and 2.306. Notice that substantial asymmetry remains even at this level of instrument strength. This is consistent with our observation in Section 6 that one-tailed t-tests using standard critical values only achieve approximate symmetry between left-tail and right tail rejection rates if first-stage F is in the tens of thousands. The ACT achieves a correct overall 5% rejection rate, as well as symmetry with 2.5% negative and 2.5% positive rejections. Recall that for a conventional two-tailed t-test these figures were 0.3% and 4.5%, and for the AR test they were 2.3% and 2.7%, respectively.

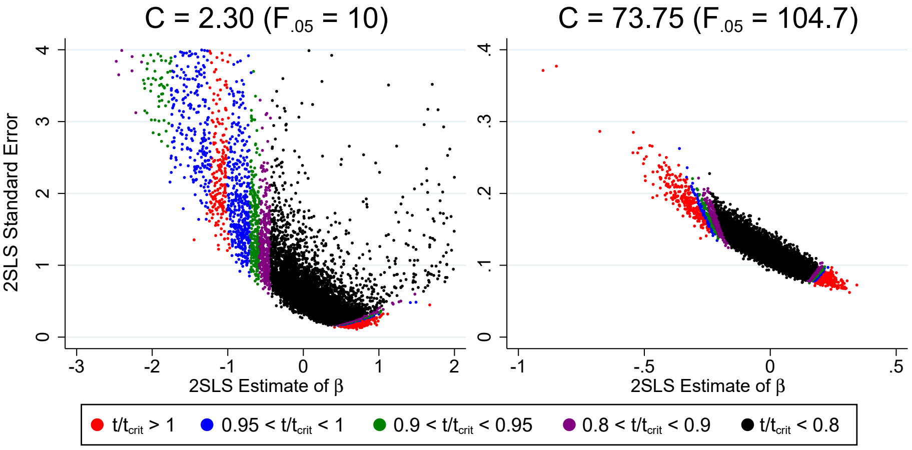
In Table 7 we examine power of the ACT test. We again consider two alternative true values, $\beta = 0.30$ or $\beta = -0.30$. These are quantitatively large values, as they imply a one standard deviation change in x induces an 0.25 standard deviation change in y.

Consider first the strong instrument case of C=74 ($F_{.05}$=105). Here the power of the ACT test is very good, with a roughly 73% rejection rate in both the $\beta = -0.3$ and $\beta = 0.3$ cases. It is interesting to compare this to the AR test, which exhibits a tremendous power asymmetry: a 91% rejection rate when $\beta = -0.3$ but only a 53.4% rejection rate when $\beta = 0.3$. This asymmetry is not specific to this example: It is a general 2SLS property that follows directly from the positive association between $\rho\widehat{cov}(z,u)$, $\hat{\beta}_{2SLS}$ and the value of the AR test (as we discussed in Section 6).

Thus, the AR test has more power to detect true effects that are opposite in direction to the OLS bias, while the ACT test has more power to detect true effects in the same

[15] We emphasize that the conditional critical values differ across the 10,000 datsets, as each dataset has its own realization of $(\hat{F}, \hat{\rho})$, but the median critical values are -0.445 and 3.152.

Figure 9. The Mills et al. (2014) Asymmetric Conditional t-test ($\rho = 0.80$)



Note: Colors signify the ratio of the t-statistic to its critical value. Runs with s.e.> 4 not shown.

direction as the OLS bias. Given that both tests have correct size by construction, the choice between them depends only on power. Suppose the researcher has a strong prior that the OLS bias is positive (i.e., positive selection into treatment). Adopting a "first do no harm" approach to policy evaluation, one would want to use the AR test, as it has more power to detect true negative effects in that context. But we would advise also implementing the ACT test, as it is more likely to detect true positive effects.

Given negative selection ($\rho < 0$) this advice is reversed, as the ACT test has more power than AR to detect a true negative effect. This may seem surprising, given the unanimous advice of the theory literature to use the AR test in the single endogenous variable just-identified case, as it is the uniformly most powerful test.[16] But that result only holds within the class of two-tailed tests with symmetric critical values.[17]

Next, we examine power of the ACT test in the case of C=2.30 ($F_{.05}$=10), often considered the standard for a "strong" instrument. In this case, power is very low. The probability of rejecting $\mathrm{H}_0 : \beta = 0$ is only 8% when true $\beta = 0.3$, and only 4.8% when true $\beta = -0.3$. In the $\beta = -0.3$ case power is no greater than size, meaning the test is not informative. Moreover, these figures are inflated by the fact that roughly 1/6 of these rejections occur when $\hat{\beta}_{2SLS}$ has the "wrong" sign. The AR test doesn't do any better. It has even lower power (only 6.6%) when true $\beta = 0.3$, and while it superficially seems to do a bit better (9%) when true $\beta = -0.3$, this is spuriously inflated by the fact that more than 1/3 of these rejections happen when $\hat{\beta}_{2SLS} > 0$. The obvious conclusion is that in the C=2.30 ($F_{.05}$=10) case there is simply not much information in the data, and no choice of testing procedure will change that.

[16] In the single endogenous variable exactly-identified case, the AR test is equivalent to the conditional likelihood ratio (CLR) test (Moreira 2003) and the Langrange multiplier (LM) test (Kleibergen 2002). In more general settings these tests differ, and the choice among them is ambiguous. Moreira (2003) argues that the power of the AR and LM tests deteriorates relative to CLR when there are many instruments.

[17] Andrews et al. (2007) found that two-tailed conditional t-tests have very poor power. In particular, when instruments are weak and $\rho > 0$, they have little power to detect a true negative β. This is because they constrain the critical values to be symmetric around zero, which fails to deal with the negative association between 2SLS estimates and standard errors that arises in the $\rho > 0$ case. The same criticism applies to the tF test proposed by Lee et al. (2020).

Table 7. Power of the ACT and AR tests ($\rho = 0.8$) (%)

C	$F_{5\%}$	Conditional t-test (ACT)			AR Test		
		H_0:$\beta = 0$	$\beta > 0$	$\beta < 0$	H_0:$\beta = 0$	$\beta > 0$	$\beta < 0$
			$\beta = 0.3$				
2.30	10	8.0	6.6	1.4	6.6	6.5	0.1
5.78	16.38	12.1	11.3	0.9	8.4	8.3	0.2
29.44	50	37.6	37.6	0.0	25.2	25.2	0.0
73.75	104.7	73.3	73.3	0.0	53.4	53.4	0.0
			$\beta = -0.3$				
2.30	10	4.8	0.7	4.2	9.0	3.2	5.9
5.78	16.38	7.3	0.4	7.0	15.0	0.8	14.2
29.44	50	38.3	0.0	38.3	54.7	0.0	54.7
73.75	104.7	73.6	0.0	73.6	91.0	0.0	91.0

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

Moving to the case of C=5.780 ($F_{.05}$=16.38) we see some improvement for both tests, as power attains levels of 7.3% to 15%, which clearly exceeds the 5% size level, and "wrong sign" rejections become rare. But these power levels still seem uninspiring. As in the strong instrument case, and for the same reason, the AR test has more power to detect a true negative β, while the ACT test has more power to detect true positives, so there is no unambiguous ranking of the tests.

Finally, we consider a moderately strong instrument case of C=29.4 ($F_{.05}$=50). At this level of instrument strength, the ACT test attains a power level of roughly 38% regardless of whether the true β is positive or negative. But the AR test still has a strong asymmetry (55% vs. 25%): It has much more power to detect a true negative β.

10. THE *T*F-TEST

Lee et al. (2020) propose a way 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 we discussed in Section 9, which conditions one-tailed t-test critical values on $\hat{F}$ and $\hat{\rho}$. The difference is that tF-test critical values are symmetric about zero, and worst-case values are assumed for both ρ and C.

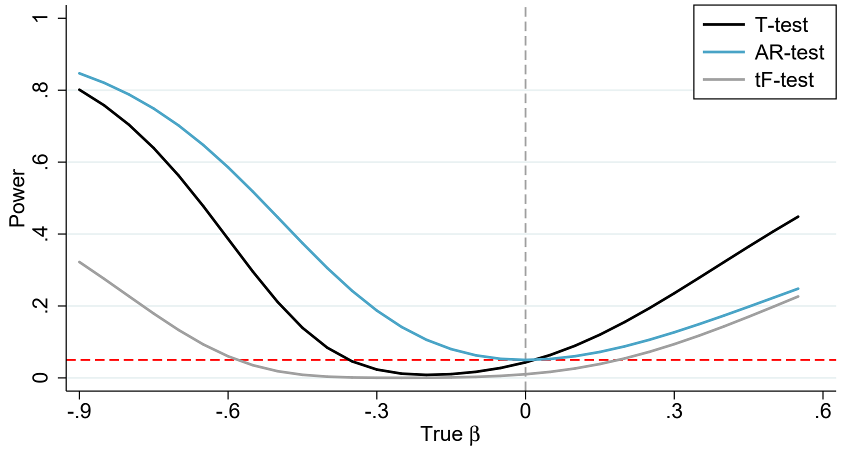
As we saw in Section 3 Figure 1, the size of the t-test is strongly increasing in ρ when instruments are weak, and the worst-case (maximal) size distortion occurs when $\rho = \pm 1$. Lee et al. (2020) show the worst case for C is $[\hat{F}/(\hat{F}^{1/2} + 1.96)]^2$. Using a procedure similar that that 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.

Lee et al. (2020) show that a first-stage sample $\hat{F}$ of at least 104.7 is required to guarantee the size of a 5% level two-tailed t-test is no greater than 5% (i.e., worst-case size distortion is zero). Hence, if the first stage $\hat{F}$ is at least 104.7 the tF test uses the conventional ±1.96 critical values to form a 5% test. 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 the AR and tF 95% confidence intervals for β are unbounded. The reason is simple: If $\hat{F} < 3.84$ then the 95% confidence interval for the first-stage coefficient π includes the case of π=0. This means we do not have 95% confidence the model is identified. It would be logically inconsistent to place a 95% confidence interval on β when we don't have 95% confidence that the model is identified. Yet a 2SLS t-test based confidence interval does exactly that when $\hat{F} < 3.84$.

By construction tF-test critical values are always greater than or equal to conventional t-test critical values.[18] So the power of the tF test is unambiguously less than that of the t-test. This can be observed in Figure 10, which compares the power curves of the tF, t and AR 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.

Figure 10. Power of the tF-test against the t-test and AR-test (C = 10, ρ = 0.5)



11. PERFORMANCE OF 2SLS RELATIVE TO OLS

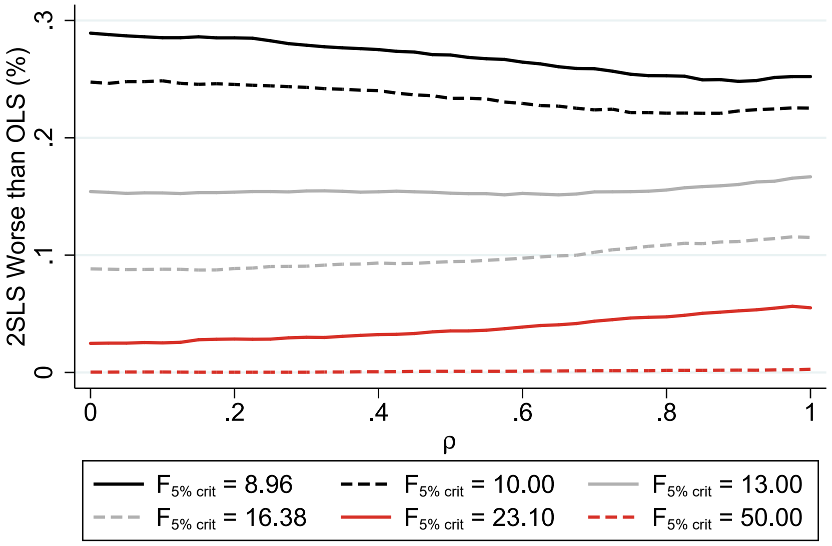
Conventional weak instrument tests ask if instruments are strong enough for 2SLS to have "nice" properties, such as a maximal (over all possible ρ) size distortion in two-tailed 2SLS t-tests below some acceptable level. But in practice applied researchers face a choice between using 2SLS and OLS. So it is also interesting to ask: "How strong do instruments need to be for 2SLS to generate more reliable results than OLS?"

We start to explore this question by asking how often 2SLS estimation errors exceed the worst-case OLS bias. This occurs if x is *perfectly* correlated with the error in the outcome equation ($\rho = 1$ and π=0). The bias is then 1.0. In Figure 11 we report the frequency of $|\hat{\beta}_{2SLS}| > 1$ for the DGP in (5). We see 2SLS estimation errors of this magnitude are common if the first-stage F is toward the low end of conventional "strong" instrument

[18] Lee et al. (2020) provide a table of tF-test 5% critical values at selected values of $\hat{F}$. To construct power curves we require a smooth function that generates the critical values as a function of $\hat{F}$. We developed the following function that approximates the critical values to high accuracy (R^2=0.9995). Letting $f = \hat{F}$ and letting $c_{.05}(f)$ denote the critical for a 5% test we have:
$c_{.05}(f) = exp(2475f^{-2} - 5709ln(f)f^{-2} + 5253ln(f)^2f^{-2} - 2395ln(f)^3f^{-2} + 543ln(f)^4f^{-2} - 50ln(f)^5f^{-2} + 0.492)$ for all F on the interval $(3.84, 104.7)$ We replace the critical values in equation (A3) with the values obtained from this function to calculate the power curve of the tF-test.

thresholds. For example, if C=2.30 ($F_{.05}$=10), which corresponds to the Staiger-Stock rule of thumb, the risk of such extraordinary large outliers is a rather remarkable 22% to 25%. Interestingly, however, if we consider a moderately strong instrument case of C=29.4 ($F_{.05}$=50) such large outliers are virtually impossible.

Figure 11. Probability of 2SLS Estimation Error Exceeding Worst-case OLS Bias.



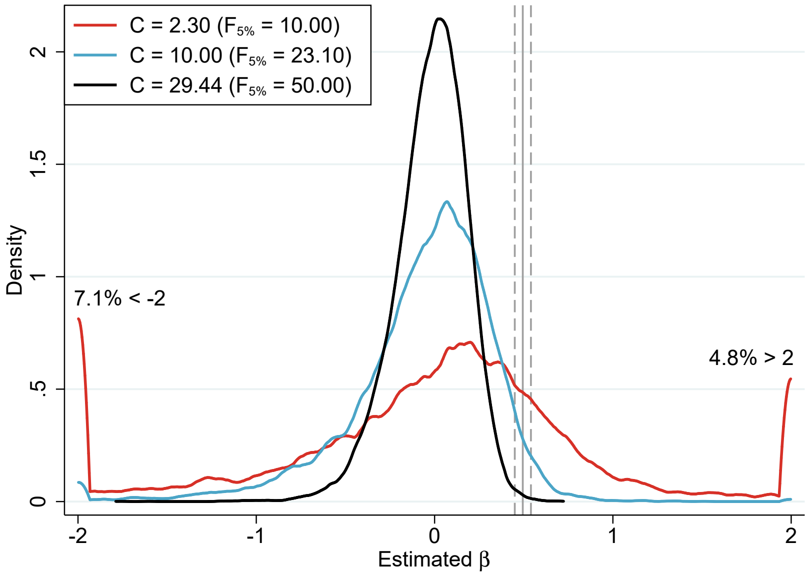
Note: We plot the proportion of times that $|\hat{\beta}_{2SLS}| > 1$, the worst-case bias of OLS.

Figure 12 compares the density of 2SLS estimates in the cases of if C=2.30 ($F_{.05}$=10), C=10 ($F_{.05}$=23.1) and C=29.4 ($F_{.05}$=50). In the first case the distribution is highly non-normal, with fat tails, high frequency of extreme outliers and left skewness very apparent. The distribution is still highly non-normal in the C=10 case. Only in the C=29.4 ($F_{.05}$=50) case does normality appear to be a decent approximation to the sampling distribution of the 2SLS estimator. Figure 12 also shows the mean OLS estimate (at 0.50) and the 95% confidence interval around that estimate. Careful inspection of the figure reveals that, due to their high dispersion, the 2SLS estimates are frequently further from the true value (β=0) than the OLS estimates. Thsi is especially true in the C=2.3, and C=10 cases, but is much less common in the C=29.4 ($F_{.05}$=50) case.

Next in Figure 13, we report the fraction of simulated datasets where 2SLS performs worse than OLS, meaning the 2SLS estimate of β is further from the truth than the OLS estimate. As expected, at low levels of endogeneity ($\rho \approx 0$) OLS is almost always better than 2SLS, as there is little bias and OLS is more efficient. What is surprising is the high frequency with which OLS outperforms 2SLS at much higher values of ρ. Take the case of C=2.30 ($F_{.05}$=10), which corresponds to the Staiger-Stock rule of thumb. The value of ρ has to approach 0.50 before the probability that 2SLS outperforms OLS passes 50%.

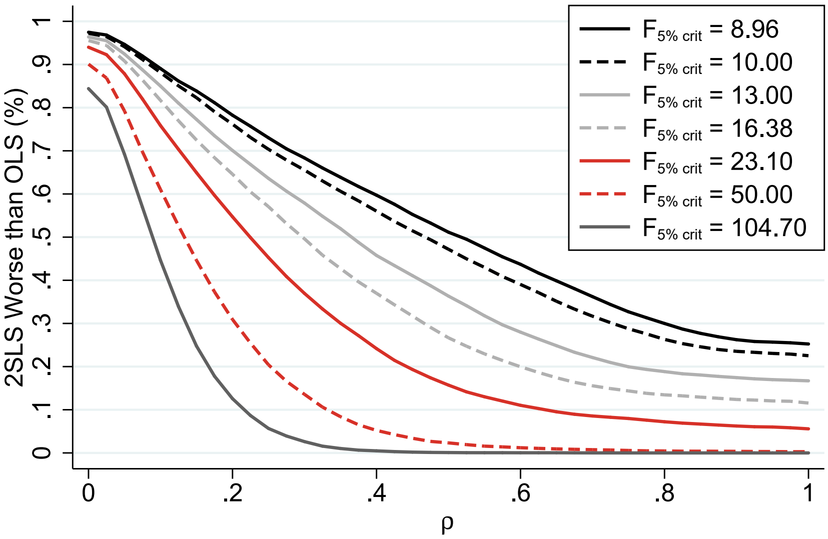
The results in Figure 13 are hard to assess without a prior on reasonable values of the unknown parameter ρ, which determines the extent of the endogeneity problem. This is not exceptional, as Stock-Yogo weak instrument tests also require an assumption on ρ. As we discussed in Section 4, those tests evaluate the *worst-case* performance of 2SLS

Figure 12. Kernel Density of 2SLS Estimates when $\rho = 0.5$



Note: This figure plots the kernel densities of 2SLS estimates censored at $+-2$ *with various values of* C*. A bandwidth of* 0.03 *was used. The grey line shows the mean OLS estimate, while the grey dotted lines represent the 95% confidence interval.*

Figure 13. Probability of 2SLS Performing Worse than OLS



Note: We plot the proportion of Monte Carlo replications where $|\hat{\beta}_{2SLS} - \beta| > |\hat{\beta}_{OLS} - \beta|$.

hypothesis tests. As we saw in Figure 1, this occurs when ρ is near one, so the endogeneity problem is very severe, so those tests are implicitly assuming this to be the case.

In many applications one can put a reasonable prior on ρ, and assess the performance of 2SLS relative to OLS for different levels of instrument strength in that scenario. For example, consider the archetypal application of IV to estimating a regression of log wages on education. Using PSID data from 2015 we calculate a correlation between education and log earnings of 0.45.[19] Thus, if education has no true effect on earnings, and the only reason it is correlated with earnings is endogeneity - i.e., it is perfectly correlated with the latent ability endowment - then the highest possible value of ρ is 0.45. So in such an application a uniform prior on $\rho \in [0, 0.45]$ may be reasonable.

Applied researchers may find a prior on ρ unfamiliar, so it is worth noting that plausible values of ρ can be backed out empirically given any hypothesized value of β. For $\beta = \beta^p$, the implied value of ρ is simply the correlation of the residuals from (i) the regression of $y - x\beta^p$ on z and (ii) the first stage regression of x on z.[20] For example, if $\beta^p = 0$ this is the correlation of the reduced form errors, and if $\beta^p = \beta_{OLS}$ this is zero. Thus, a prior that $\beta \in (0, \beta_{OLS})$, which may be natural in many applications where one suspects positive selection into treatment, would correspond to a prior that ρ lies between zero and the correlation of the reduced form residuals. This is how we motivate the uniform prior on $\rho \in [0, 0.45]$ in the example of wages and education.

Table 8 shows the probability that 2SLS will outperform OLS under different scenarios for the concentration parameter C (and the associated first-stage F), and different priors on ρ. For example, if C=2.30 ($F_{.05}$=10), which corresponds to the Staiger-Stock rule of thumb for strong instruments, and given a uniform prior $\rho \in [0, 0.45]$, the probability that 2SLS outperforms OLS is only 26%. Alternatively, a researcher who thinks education is highly (but not completely) endogenous, might have a uniform prior of $\rho \in [0.35, 0.45]$. Even in that case, the probability that 2SLS outperforms OLS is only 41%.

Clearly then, in an application to estimating the effect of education on earnings, one should require a substantially higher level of instrument strength than the $\hat{F}$=10 threshold suggests. For instance, in the case of C=29.4 ($F_{.05}$=50), a uniform prior on $\rho \in [0, 0.45]$ implies a 65% chance that 2SLS will outperform OLS. Given a uniform prior on $\rho \in [0.35, 0.45]$ this increases 95%. Thus, in the archetypal application of IV to estimating the return to education, if one believes ability bias is severe, one needs an $\hat{F}$ in the vicinity of at least 50 to have high confidence that 2SLS will outperform OLS.

Of course in other contexts the endogeneity problem is plausibly more severe. For example, consider a case where a uniform prior of $\rho \in [0.5, 1.0]$ is plausible. Even then, if C=2.30 ($F_{.05}$=10), which corresponds to the Staiger-Stock rule of thumb for strong instruments, the probability that 2SLS outperforms OLS is only 69%. But in the moderately strong instrument case of C=10.0 ($F_{.05}$=23.1) the chance of 2SLS outperforming OLS is 91%. One may contrast this with the case of a uniform prior $\rho \in [0, 0.45]$, where this level of instrument strength only gives a 47% chance that 2SLS will outperform OLS. Thus, an "acceptable" level of instrument strength such that 2SLS is likely to outperform OLS clearly depends heavily on one's prior about ρ.

Overall, the results of this section show instruments need to be much stronger than standard thresholds, like the popular $\hat{F} > 10$, in order to (i) assure 2SLS does not suffer from a high probability of extreme outliers, and (ii) give confidence that 2SLS results are superior to OLS, in the sense that $|\hat{\beta}_{2SLS} - \beta| < |\hat{\beta}_{OLS} - \beta|$. It is difficult to

[19] We use data on 30-54 year old household heads, and we partial out effects of age and age^2. The wage is constructed as labor income/hours. We screen on hours $\in [400, 4160]$, income $\in [\$3000, \$235884]$, wage>\$2.70 per hour, and valid data on education and labor income. This gives N=3,634.

[20] Of course also including any exogenous control variables present in the application.

give any general rule of thumb for acceptable instrument strength by the latter metric, as the probability that 2SLS will outperform OLS is strongly increasing in the degree of endogeneity (ρ). We suggest that researchers assess the level of instrument strength required to have reasonable confidence that 2SLS will outperform OLS in any particular application, based on reasonable priors on the severity of the endogeneity problem (ρ).

Nevertheless, a strong case can be made that applied researchers should adopt a higher threshold of instrument strength in the vicinity of $\hat{F} > 50$. As we have seen, this threshold renders extreme outlier 2SLS estimates very unlikely, and it makes 2SLS likely to outperform OLS even at moderate levels of ρ. And, as we saw in Section 9, the power of robust tests (AR and ACT) is quite poor if instrument strength falls much below this level. If such a threshold cannot be met, then OLS combined with a serious attempt to control for sources of endogeneity may be a superior strategy in many cases. We reiterate that robust tests should be used in lieu of 2SLS t-tests regardless of $\hat{F}$.

Table 8. Probability of 2SLS Outperforming OLS (%)

Concentration	$F_{5\%Crit}$	Uniform Prior for ρ:			
Parameter		0 to 1	0 to 0.45	0.35 to 0.45	0.5 to 1
1.82	8.96	45	24	41	65
2.30	10	48	26	45	69
3.84	13	56	32	55	77
5.78	16.38	62	38	64	84
10.00	23.1	70	47	77	91
29.44	50	83	65	95	99
73.75	104.7	89	76	100	100

Note: We report the frequency of $|\hat{\beta}_{2SLS} - \beta| < |\hat{\beta}_{OLS} - \beta|$ across Monte Carlo replications, averaged across all possible values of ρ under a uniform prior that ρ falls in the indicated range.

Finally, we note that our whole discussion of properties of 2SLS has been centered on the *iid* normal case, in order to focus on key issues. This is not as restrictive as it may appear, as for any heteroskedastic DGP, there exists a homoskedastic DGP yielding equivalent behavior for 2SLS estimates and test statistics - see Andrews et al. (2019). But in assessing acceptable first-stage F statistics in practice it is important to consider the impact of heteroskedasticity. In the *general* case of multiple instruments, as Andrews et al. (2019) note, it is inappropriate to use either a conventional or heteroskedasticity robust F-test to gauge instrument strength in non-homoskedastic settings. They suggest using the Olea and Pflueger (2013) effective first-stage F-statistic. However, as they point out, in the single instrument just-identified case, this reduces to the conventional robust F, and also coincides with the Kleibergen and Paap (2006) Wald statistic.

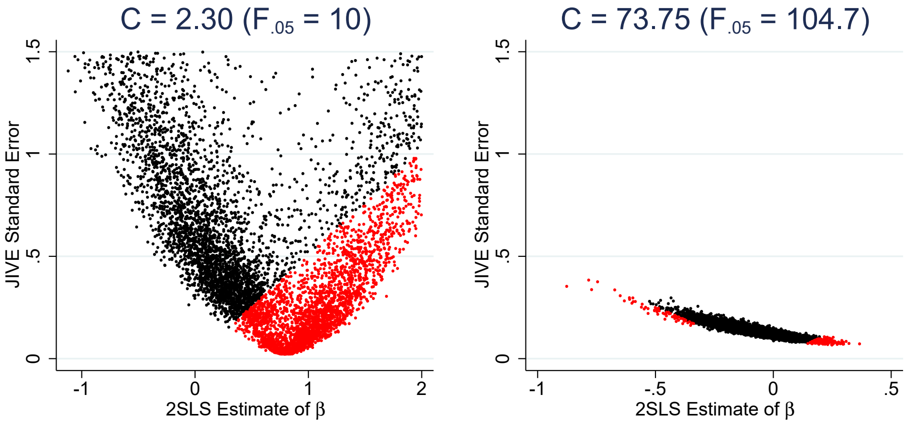
12. IS THERE A BETTER ALTERNATIVE TO 2SLS?

We have found that 2SLS performs very poorly in the $F_{.05}$=10 case often viewed as a benchmark for acceptably strong instruments. It not only exhibits poor size and power properties, but in many plausible cases it is likely to underperform OLS, in that $\hat{\beta}_{2SLS}$ is likely to be further from the true value than $\hat{\beta}_{OLS}$. In this Section we consider three alternatives to 2SLS and ask whether they perform better when instruments are weak.

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_ie_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.[21]

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 report results using JIVE in Figure 14.

Figure 14. Standard Error of $\hat{\beta}_{JIVE}$ plotted against $\hat{\beta}_{JIVE}$ itself ($\rho = 0.80$)



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

In the case of C=2.30 ($F_{.05}$=10) the JIVE estimator causes us to reject H$_0$: $\beta = 0$ via a two-tailed 5% t-test a striking 29% of the time, and *all* the rejections are positive. In Sections 4-6 we emphasized the problem that 2SLS is much more likely to judge estimates significant if they are shifted in the direct of the OLS bias. Here we see that JIVE makes this problem much worse. The negative association between $se(\hat{\beta}_{JIVE})$ and $\hat{\beta}_{JIVE}$ imparts positive $\hat{\beta}_{JIVE}$ estimates with spuriously high precision.

JIVE performs worse than 2SLS 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.[22] In the right panel of Figure 14 we see that in the relatively strong instrument case of C=73.75 ($F_{.05}$=104.7)

[21] 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. Similarly, if instruments are strong in the sense discussed in Section 2, so we can be confident that $|\pi\widehat{Var}(z)| \gg |\widehat{cov}(z,e)|$, then the influence of any particular e_i on $\hat{\pi}$ becomes negligible.

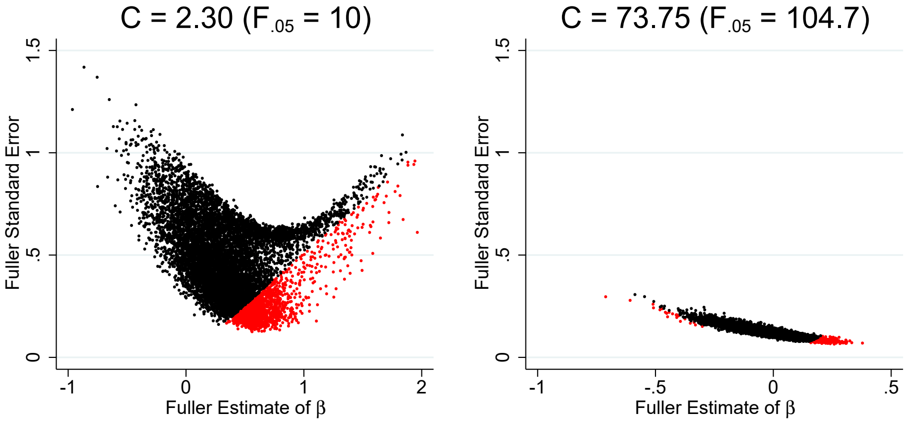
[22] In fact, in the runs in the left panel of Figure 14, $\widehat{cov}(z,x)$ is always positive, as we would hope given that $cov(z,x) > 0$ in the population. But $\widehat{cov}(z\hat{\pi}_{-i}, x)$ has an incorrect negative sign 30% of the time!

JIVE does somewhat better (i.e, 3.84% rejections of which 83% are positive), but this is not comforting for an estimator designed for use with weak instruments.

The "k-class" estimators modify 2SLS by implementing IV using $kz_i\hat{\pi} + (1-k)x_i$ as the instrument for x_i. Obviously 2SLS uses $k = 1$. One important alternative to 2SLS is the Fuller (1977) estimator that uses $k = 1 - 1/N$ (in the one instrument case), thus leaving in a small part of x_i. This "stabilizes" the estimator. Hence, in contrast to 2SLS, the mean and variance of Fuller's estimator exist.

We report results using the Fuller estimator in Figure 15. First, consider the case of $\rho = 0.80$ and C=2.30 ($F_{.05}$=10). Comparison with Figure 2 shows that Fuller estimates and standard errors are substantially less dispersed than 2SLS. Fuller causes us to reject H$_0$: $\beta = 0$ via a two-tailed 5% t-test 15.4% of the time, compared to 10% for 2SLS, so its size distortion is greater. Just as with 2SLS, *all* the rejections occur when $\hat{\beta}_{Full}$ is positive: the negative covariance between $se(\hat{\beta}_{Full})$ and $\hat{\beta}_{Full}$ imparts spuriously high precision to Fuller estimates that are most shifted in the direction of the OLS bias.

Figure 15. Standard Error of $\hat{\beta}_{Full}$ plotted against $\hat{\beta}_{Full}$ itself ($\rho = 0.80$)



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

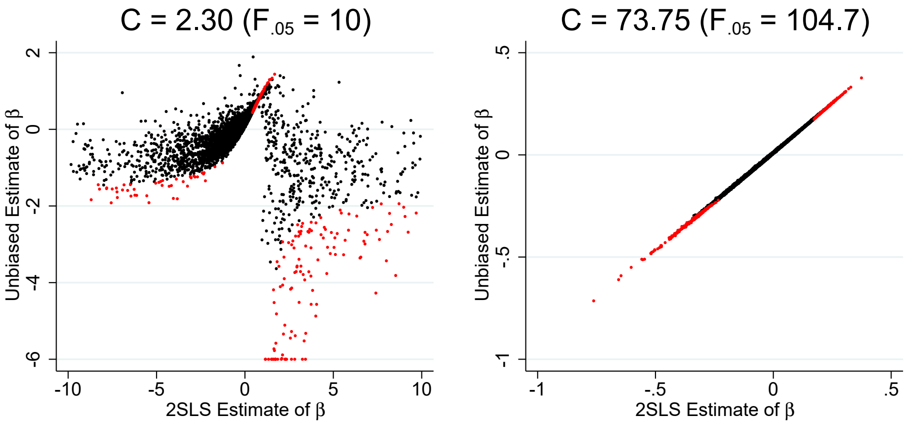
In the right panel Figure 15 we report results for the relatively strong instrument case of C=73.75 ($F_{.05}$=104.7). Comparison with the right panel of Figure 2 reveals that 2SLS and Fuller estimates are very similar in this case.[23]

The third and final alternative we consider is the unbiased estimator of β proposed in Andrews and Armstrong (2017). To understand their approach, consider the reduced form regression of y on z, $y = z\beta\pi + (\beta e + u) = z\xi + v$ where $\xi = \beta\pi$. Obviously $\beta = \xi/\pi$. By analogy, the 2SLS estimator can be calculated by taking the ratio $\hat{\beta}_{2SLS} = \hat{\xi}/\hat{\pi}$ where $\hat{\pi}$ is the first-stage estimate of π. The problematic properties of 2SLS that arise when instruments are weak (see Section 2) may be understood as arising because $\hat{\pi}$ appears in the denominator of this ratio. Estimates of ratios have poor properties if the denominator is noisy, including the fact that $1/\hat{\pi}$ is not an unbiased estimator of $1/\pi$.

[23] The Fuller estimator rejects at a 5.43% rate, but 96% of the rejections are when $\hat{\beta}_{Full} > 0$. So just as with 2SLS, severe one-tailed t-test size distortions persist even with quite strong instruments.

The Andrews-Armstrong idea is that an unbiased estimator of β can be obtained if one can construct an unbiased estimator of $1/\pi$. Their approach requires the researcher to be certain that $\pi > 0$, which is plausible in many applications. In that case, an unbiased estimate of $1/\pi$ can be constructed by taking $1/\hat{\pi}^*$, where $\hat{\pi}^* = \sigma_2\phi(\hat{\pi}/\sigma_2)/(1-\Phi(\hat{\pi}/\sigma_2))$. Here σ_2 is the standard deviation of $\hat{\pi}$, and ϕ and Φ are the standard normal density and cdf. Given $\hat{\pi}^*$ it is simple to construct an unbiased estimator that we denote $\hat{\beta}_U$.[24]

Figure 16. Andrews-Armstrong $\hat{\beta}_U$ plotted against $\hat{\beta}_{2SLS}$ ($\rho = 0.80$)



Note: Runs where $\hat{\beta}_U < -6$ were censored to -6. Red dots indicate $H_0 : \beta = 0$ is rejected at the 5% level using the Anderson-Rubin statistic.

It is important to understand how the first-stage estimate $\hat{\pi}$ is modified by this transformation. Note $\phi/(1-\Phi)$ is the inverse Mills ratio, so $\hat{\pi}^* = E(x|x > \hat{\pi})$ where $x \sim N(0, \sigma_2^2)$. Thus $\hat{\pi}^*$ is positive by construction, and always larger than $\hat{\pi}$. If $\hat{\pi}$ is negative, then $\hat{\pi}^*$ is a small positive number. As $\hat{\pi}$ grows large $\hat{\pi}^*$ approaches $\hat{\pi}$ from above.

We report results using $\hat{\beta}_U$ in Figure 16, where we plot the $\hat{\beta}_U$ against $\hat{\beta}_{2SLS}$. The red dots indicate estimates that are significant at the 5% level according to the Anderson-Rubin test. In the strong instrument case in the right panel, $\hat{\beta}_U$ and $\hat{\beta}_{2SLS}$ are nearly identical. The AR test rejects H$_0$: $\beta = 0$ at close to the correct 5% rate, and there is an fairly even balance between rejections at positive vs. negative estimates of β.

The weak instrument case in the left panel is more interesting. Of course the AR test rejects H$_0$: $\beta = 0$ at close to the correct 5% rate. But for 2SLS the rejections are highly asymmetric, as 85% occur when $\hat{\beta}_{2SLS} > 0$. So using AR does not avoid the asymmetry that most rejections occur at positive values, which is the direction of the OLS bias. As we see in the left panel of Figure 16, the unbiased estimator solves this problem, as it generates only 54% positive rejections. It achieves this in an interesting way: Specifically, it flips a large fraction of the significant 2SLS estimates from positive to negative. This occurs in cases where $\hat{\pi}$ is negative, so that $\hat{\pi}^*$ is positive (consistent with the prior).

[24] The unbiased estimator of β is simply $\hat{\beta}_U = (\hat{\delta}/\hat{\pi}^*)+(\sigma_{12}/\sigma_2^2)$, where $\hat{\delta}$ is defined as $\hat{\xi}-(\sigma_{12}/\sigma_2^2)\hat{\pi}$. This works because $E(\hat{\xi}|\hat{\pi}) = \beta\pi + (\sigma_{12}/\sigma_2^2)(\hat{\pi} - \pi)$ where σ_{12} is the covariance between $\hat{\pi}$ and $\hat{\xi}$. Therefore $E(\hat{\delta}|\hat{\pi}) = \beta\pi - (\sigma_{12}/\sigma_2^2)\pi$, which is independent of $\hat{\pi}$. Hence we can write $E(\hat{\delta}/\hat{\pi}^*) = \beta - (\sigma_{12}/\sigma_2^2)$, from which it follows that $E\hat{\beta}_U = \beta$. To obtain a feasible estimator replace σ_{12} and σ_2 with their estimates.

Finally, Table 9 repeats the analysis of Section 11, by asking how often the alternative estimators outperform OLS, in the sense that $|\hat{\beta} - \beta| < |\hat{\beta}_{OLS} - \beta|$. In the $F_{.05} = 50$ case the Fuller and Unbiased estimators perform about the same as 2SLS, with JIVE slightly worse, so at this level of instrument strength there is little to be gained by using alternatives to 2SLS. When instruments are weaker ($F_{.05} = 10$ or 23) a clear ranking is evident with Fuller doing best, followed by Unbiased, then 2SLS and then JIVE. For instance, given a uniform prior $\rho \in [0, 0.45]$, which we have argued is plausible in the classic application of estimating returns to education, the probability that 2SLS outperforms OLS is only 26% when $F_{.05} = 10$. For Fuller the figure is 40% and for Unbiased it is 32%. So while these alternatives outperform 2SLS, their performance can hardly be considered acceptable in any absolute sense when instruments are weak.

Table 9. Probability of Estimators Outperforming OLS (%)

Estimator	Prior Expectation of ρ:			
	0 to 1	0 to 0.45	0.35 to 0.45	0.5 to 1
	$C = 2.30$, $F_{5\%Crit} = 10.00$			
2SLS	48	26	45	69
JIVE	32	22	33	41
Fuller	65	40	65	87
Unbiased	60	32	56	84
	$C = 10.00$, $F_{5\%Crit} = 23.10$			
2SLS	70	47	77	91
JIVE	57	38	62	74
Fuller	76	51	84	98
Unbiased	75	50	82	96
	$C = 29.44$, $F_{5\%Crit} = 50.00$			
2SLS	80	60	92	98
JIVE	76	56	86	93
Fuller	82	62	94	99
Unbiased	82	62	94	99

Note: We report the frequency of $|\hat{\beta} - \beta| < |\hat{\beta}_{OLS} - \beta|$ across Monte Carlo replications, averaged across all possible values of ρ under a uniform prior that ρ falls in the indicated range.

The story changes, however, if the endogeneity problem is more severe. Given a uniform prior $\rho \in [0.5, 1.0]$ the probability that 2SLS outperforms OLS is only 69% when $F_{.05} = 10$. For Fuller the figure is 87% and for Unbiased it is 84%. So these alternatives do offer substantially improved performance over 2SLS when endogeneity is severe.

In summary, these results reinforce our earlier conclusion that a first-stage F of at least 50 is required to give reasonable confidence that any of the IV estimators will outperform OLS at moderate levels of ρ. But these estimators do offer improvements in cases where instruments are weaker and endogeneity is severe. All of these alternative estimators should be used in conjunction with the AR and ACT tests.

We have focused on the one instrument case, but the performance of 2SLS tends to deteriorate with multiple instruments – see Section 13. The absolute performance of the alternatives (LIML, Fuller, JIVE, Unbiased) will also deteriorate, but less so. Hence, an even higher threshold of instrument relevance is desirable with multiple instruments.

13. THE CASE OF MULTIPLE INSTRUMENTS

Finally, we consider the over-identified case with one endogenous variable and multiple instruments. In the interest of space we focus on the case of three instruments, the number required for the mean and variance of the 2SLS estimator to exist.

Given K instruments the definition of the concentration parameter C is unchanged. But "true F" is now C/K and the first-stage sample $\hat{F}$ is $(N/K)\hat{Var}(z\pi)/\hat{\sigma}_e^2$ which has an $F(K, N-1)$ distribution. Table 10 lists, in the $K = 3$ case, several different levels of C, the associated true F, and the associated 5% critical values of the $F(3, \infty)$ distribution.

We continue to work with the model in equation (5), 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 10 then lists the value of π required to attain each level of the concentration parameter when N=1000.

Table 10. First Stage F Critical Values Required to Achieve Different Objectives

Concentration Parameter	True First-Stage F Statistic	Value of π	$F_{5\%}$	Goal Achieved
6.90	2.30	0.0480	6.93	Matching π
13.01	4.34	0.0659	**10.00**	SS Rule of Thumb
40.91	13.64	0.1168	22.30	Size = 10%
110.55	36.85	0.1920	50.00	
360.26	120.09	0.3465	142.50	Size = 5%

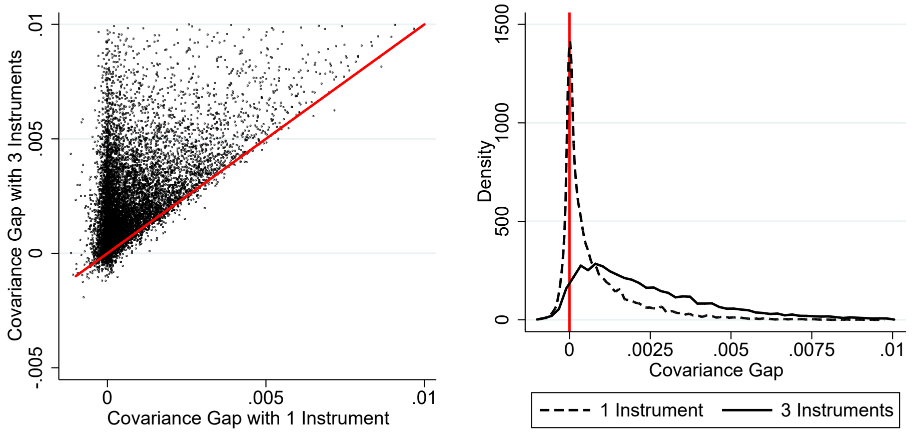
Note: The instrument vector z is significant at the 5% level in the first stage if F>2.60.

Recall from Table 1 that setting π=0.048 in the one instrument case generates a concentration parameter and "true F" of 2.3, and an associated 5% critical value of 10 for the $F(1, \infty)$ distribution. Suppose now we have three equally strong instruments. As we see in Table 10 this triples C to 6.9, leaves "true F" unchanged at 2.3, and now the associated 5% critical value of the $F(3, \infty)$ distribution is 6.93.

One might think that three equally strong independent instruments are better than one, but it is actually a mixed bag. Recall 2SLS is IV using $z\hat{\pi}$ as the instrument for x, where $\hat{\pi}$ is obtained from OLS regression of x on z. Many problematic properties of 2SLS arise because the sample covariance of the feasible instrument $z\hat{\pi}$ with the structural error u tends to be greater than that of the optimal instrument $z\pi$, by virtue of how OLS forms $\hat{\pi}$. And, unfortunately, the use of multiple instruments in the first stage of 2SLS tends to worsen the problem. This is because an instrument that happens to have a high sample covariance $\hat{cov}(z, e)$ with the first-stage error e will get a larger coefficient $\hat{\pi}$ in the first stage. This drives up the sample covariance $\hat{cov}(z\hat{\pi}, e)$. And if e and u are correlated (i.e., if we have endogeneity) this also drives up the magnitude of $\hat{cov}(z\hat{\pi}, u)$.

Figure 17 illustrates this problem. We first define the "covariance gap" as the difference $\hat{cov}(z\hat{\pi}, u) - \hat{cov}(z\pi, u)$. Using 10,000 artificial data sets generated from model (5) with ρ = 0.5 and N=1000, we calculate this covariance gap for the cases of 3 vs. 1 instrument. The left panel of Figure 13 shows how the covariance gap almost always increases, and often substantially, in the three instrument case. The right panel shows that the density of the covariance gap shifts sharply to the right. Thus, using three instruments greatly increases the problem of sample covariance between the instruments and the structural error, which will tend to bias the 2SLS estimate towards OLS.

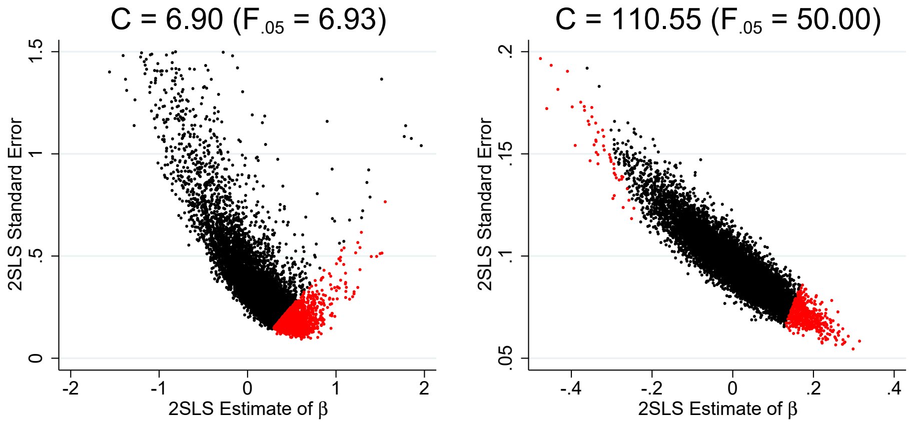
Figure 17. Instrument Endogeneity in Samples with One vs. Three Instruments



Note: We construct the "covariance gap" $\hat{cov}(z\hat{\pi}, u) - \hat{cov}(z\pi, u)$ for cases of 1 and 3 instruments with $\pi = 0.048$. We plot their joint distribution (left), and their marginal densities (right).

As a consequence of increased sample covariance $\hat{cov}(z\hat{\pi}, u)$, the association between 2SLS estimates and their standard errors gets even stronger in the multiple instrument case. This is shown in Figure 18. The left panel is comparable in all respects to the left panel of Figure 2, except now we add two equally strong independent instruments so K=3. This causes the Spearman r_s to increase in magnitude from -.576 to -.781. As the variance of the 2SLS estimator now exists, we can report a Pearson correlation of -0.547.

Figure 18. $se(\hat{\beta}_{2SLS})$ plotted against $\hat{\beta}_{2SLS}$ with Three Instruments ($\rho = 0.80$)


C = 6.90 (F.05 = 6.93)

C = 110.55 (F.05 = 50.00)

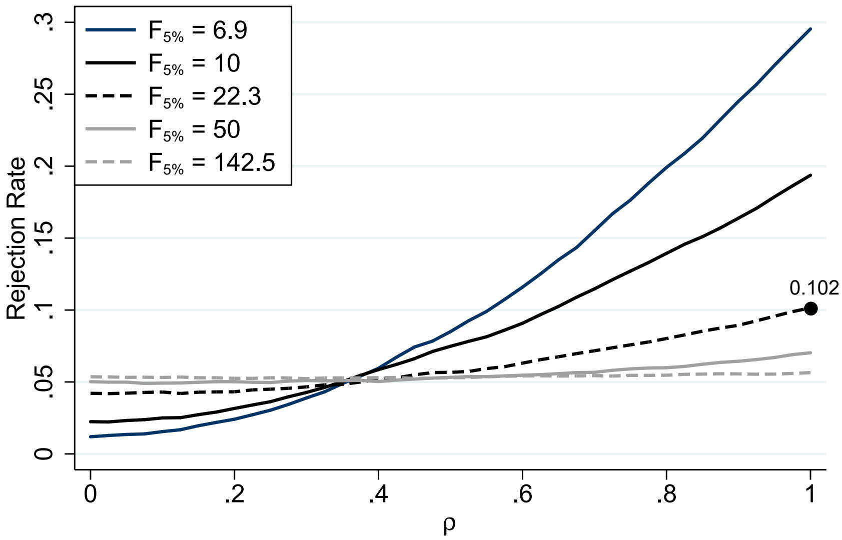
Note: Runs with std. error > 1.5 not shown. Red dots indicate $H_0 : \beta = 0$ rejected at 5% level.

By comparing the left panels of Figures 2 and 18 we can see how adding two instruments affects size of the 2SLS t-test. The red dots again indicate cases where $\hat{\beta}_{2SLS}$ differs significantly from zero according to a two-tailed 5% t-test. With one instrument the size of the test was 10%, but now it is 19.9%, so the size distortion increases dramatically. As before, <u>all</u> rejections occur when $\hat{\beta}_{2SLS} > 0$. Due to covariance between 2SLS estimates and standard errors, only estimates shifted towards the OLS bias are ever significant.

The right panel of Figure 18 reports results for a strong instrument case of C=110.6 ($F_{.05}$=50). Here the the Spearman r_s between the 2SLS estimates and their standard errors is -.906, while the Pearson correlation is -.915. This illustrates our point from Section 5 that this strong association persists in strongly identified models. This, in turn, explains why 2SLS t-tests are unreliable even with strong identification. The 5% rejection rate of the two-tailed t-test is now 6%, so the size distortion is mostly eliminated. But fully 92% of those rejections occur when $\hat{\beta}_{2SLS}$ >0, so the asymmetry is still severe.

Figure 19 plots how size of two-tailed 5% t-tests of H$_0$:$\beta = 0$ depend on C and ρ. It is interesting to compare the case of C=2.3 in Figure 1 with C=6.9 in Figure 19. This corresponds to adding two new independent instruments of equal strength. Notice how in the 3 instrument case the rejection rate rises much more steeply with ρ, and it peaks at almost 30%, compared to only 13% in the one instrument case. Adding instruments worsens t-test size distortions by increasing the sample covariance between $z\hat{\pi}$ and u.

Figure 19. Rejection rate of H$_0$: $\beta = 0$ using a 5% Two-tailed t-test (3 Instruments)

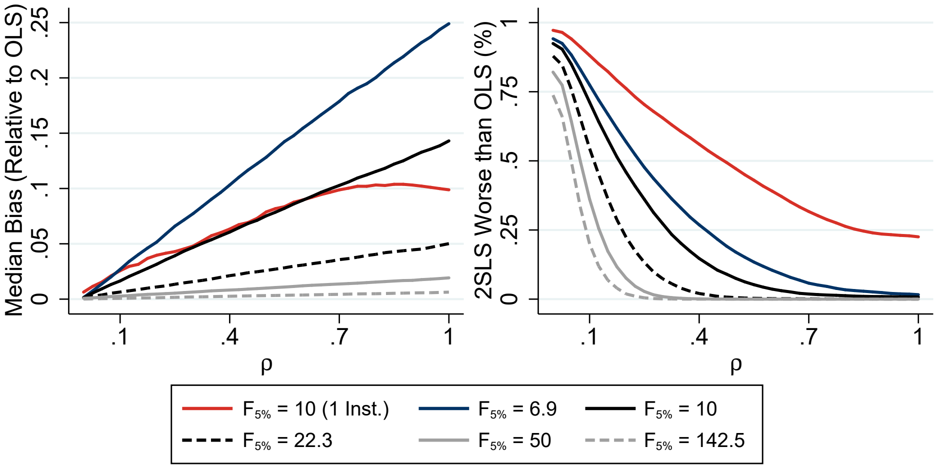


To achieve the Staiger-Stock rule of thumb ($F_{.05}$=10) in the three instrument case we need to increase π from 0.0480 to 0.0659, so we need three independent instruments that are each individually stronger than what we would require of a single instrument to achieve the same goal. Even then, as we see in Figure 19, the maximum rejection rate increases to almost 20%, compared to 13% in the single instrument case.

Stock and Yogo (2005) show that C=40.9 ($F_{.05}$=22.3) achieves a maximum rejection rate of 10% in the three instrument case. Figure 19 shows this is accurate. This corresponds to a π of 0.1168 on each of the three instruments in the first stage (see Table 10). But as we saw in Table 1, the same objective could be achieved using a single instrument with π=0.0760. Thus, if size distortion 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.

A similar point emerges if we look at median bias. The left panel of Figure 20 shows how median bias varies with C and ρ in the three instrument case. For comparison, the red line shows the case of one instrument with $C = 2.3$ ($F_{.05}$=10). This can be compared to the blue line, which is the case where we add two additional equally strong independent instruments. Clearly this makes the median bias unambiguously worse.

Figure 20. Performance of 2SLS Relative to OLS



Note: We plot median bias and proportion of Monte Carlo runs where $|\hat{\beta}_{2SLS} - \beta| > |\hat{\beta}_{OLS} - \beta|$.

It would be a mistake, however, to focus exclusively on median bias and size distortions of t-tests in assessing the performance of 2SLS. Efficiency and power considerations are also important, and robust tests are available. To explore efficiency, the right panel of Figure 20 shows how the probability of 2SLS performing worse than OLS varies with C ($F_{.05}$) and ρ in the three instrument case. We plot the proportion of simulated datasets where $|\hat{\beta}_{2SLS} - \beta| > |\hat{\beta}_{OLS} - \beta|$. For comparison, we also show (in red) the case of one instrument with $C = 2.3$($F_{.05}$=10). This may be usefully compared to the blue line, which is the case where we add two more equally strong independent instruments.

The addition of two equally strong independent instruments tremendously increases the probability that 2SLS will out-perform 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.

These results naturally lead us to assess whether robust tests outperform the t-test in the multiple instrument case. Table 11 compares the t-test, the Anderson and Rubin (1949) test and Moreira (2003)'s conditional likelihood ratio test (CLR) in the three instrument case. We focus on the case of ρ=0.80, and let the true β be 0, -0.3 or +0.3. As we noted in Section 4, $\beta = \pm 0.3$ correspond to fairly large effects, as they imply a one standard deviation change in x induces an 0.25 standard deviation change in y.

First consider the t-test. The top panel of Table 11 reports results when β=0. Clearly, the two-tailed t-test rejects the true null at far too high a rate in cases with C=6.9 ($F_{.05}$=6.93) or C=13 ($F_{.05}$=10). The size distortion in one-tailed t-tests is even greater, as all rejections occur when $\hat{\beta}_{2SLS} > 0$. In the middle panel, when β=-0.30, we see the t-test has essentially no power to detect a substantial true negative effect in these cases.

Stunningly, if C=6.9, the t-test rejects H$_0$:β=0 only 2.5% of the time, and <u>all</u> of those rejections happen when $\hat{\beta}_{2SLS} > 0$ – i.e., we conclude β is positive when it is actually negative! Lastly, consider the bottom panel, where β=0.30. In the C=6.9 case the t-test rejects H$_0$:β=0 at a 46% rate. But this high value is not a good sign – it arises because the 2SLS standard errors are spuriously precise when $\hat{\beta}_{2SLS} > 0$.

Table 11. Rejection Rates and Power With Three Instruments ($\rho = 0.8$) (%)

C	6.90	13.01	40.91	110.55	360.26
$F_{5\%Crit}$	6.93	10.00	22.30	50.00	142.50
$\beta = 0$					
T-Statistic					
Total Rejection Rate	0.198	0.134	0.082	0.060	0.054
When $\hat{\beta} > 0$	0.198	0.134	0.082	0.055	0.040
AR Test					
Total Rejection Rate	0.051	0.051	0.051	0.051	0.051
When $\hat{\beta} > 0$	0.048	0.042	0.036	0.032	0.029
CLR Test					
Total Rejection Rate	0.050	0.049	0.049	0.048	0.049
When $\hat{\beta} > 0$	0.036	0.028	0.023	0.023	0.024
$\beta = -0.3$					
T-Statistic					
Total Rejection Rate	0.025	0.007	0.354	0.949	1.000
When $\hat{\beta} < 0$	0.000	0.001	0.354	0.949	1.000
AR Test					
Total Rejection Rate	0.120	0.189	0.521	0.936	1.000
When $\hat{\beta} < 0$	0.074	0.166	0.519	0.936	1.000
CLR Test					
Total Rejection Rate	0.162	0.268	0.683	0.980	1.000
When $\hat{\beta} < 0$	0.140	0.263	0.683	0.980	1.000
$\beta = 0.3$					
T-Statistic					
Total Rejection Rate	0.460	0.453	0.583	0.838	0.997
AR Test					
Total Rejection Rate	0.075	0.097	0.221	0.540	0.977
CLR Test					
Total Rejection Rate	0.097	0.138	0.327	0.708	0.993

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

Next we consider the AR test. It performs much better than the t-test, but it still exhibits serious problems. When the true β=0 (top panel) the AR test of course rejects H$_0$:β=0 at approximately the correct 5% rate (deviating only due to sampling variation). But in the C=6.9 ($F_{.05}$=6.93) and C=13 ($F_{.05}$=10) cases nearly all of those rejections occur when $\hat{\beta}_{2SLS} > 0$. With one instrument we found this power asymmetry vanished very quickly as instrument strength increased, but in the three instrument case it vanishes more slowly. For instance, even in the strong instrument case of C=111 ($F_{.05}$=50) we see that 61% of the AR test rejections occur when $\hat{\beta}_{2SLS} > 0$.

The AR test assigns spurious precision to 2SLS estimates shifted in the direction of the OLS bias, and in the multiple instrument case this problem is much worse. This problem arises due to the strong positive covariance between $\rho\widehat{cov}(z,u)$ and $\hat{\beta}_{2SLS}$. A large value of $\rho\widehat{cov}(z,u)$ also generates a large value of the AR test. Thus, if $\rho > 0$, the AR test and $\hat{\beta}_{2SLS}$ have a positive covariance. Adding instruments increases this covariance.

The problem with the AR test is even more apparent if we look at the case where the true β=-0.3 (middle panel). In the case of C=6.9 ($F_{.05}$=6.93) the AR test rejects the false null H_0:β=0 at a 12% rate, but a substantial fraction of those rejections (39%) occur when $\hat{\beta}_{2SLS} > 0$. So we often conclude β is positive when it is actually negative.

As we discussed in Section 6, the AR test has more power to detect true negative effects than true positive effects. This is another manifestation of the covariance phenomenon. In the strong instrument case of C=111 ($F_{.05}$=50) the AR test rejects H_0:β=0 at a 93.6% rate when the true β is negative, but only a 54% rate when the true β is positive. Geometrically, when the true beta is positive, $\hat{\beta}_{2SLS}$ values that lie between zero and the true β (against the direction of the OLS bias) will correspond to spuriously low AR test values, so they are less likely to be judged significantly greater than zero.

Next we consider the CLR test. Its performance is clearly superior to AR. When the true β=0 (top panel) the CLR test of course rejects H_0:β=0 at approximately the correct 5% rate. In the C=6.9 ($F_{.05}$=6.93) case 73% of those rejections occur when $\hat{\beta}_{2SLS} > 0$, so the CLR test does exhibit power asymmetry. But this vanishes rather quickly as instrument strength increases. In the C=13 ($F_{.05}$=10) case the proportion of positive rejections is already down to 57%. Like AR, the CLR test has more power to detect true negative effects than true positive effects, but the asymmetry is less severe. For example, In the case of C=41 ($F_{.05}$=22.3) the CLR test rejects H_0:β=0 at a 68.3% rate when the true β is negative, and a 32.7% rate when the true β is positive. These compare to rates of 52.1% and 22.1% for the AR test. Thus, the CLR test exhibits better power to detect both positive and negative departures from the null.

Finally, consider the ACT test; see Table 12. In contrast to the other tests, if β=0 the ACT test has correct 5% size and symmetric rejections on both sides of 0. Moreover, the ACT test has roughly symmetric power when the true $\beta = \pm 0.30$. Interestingly, the ACT test has better power to detect a true positive β (when the OLS bias is positive) than the CLR test. Conversely, CLR has better power to detect a true negative effect. This is consistent with results in Table 7 for the exactly identified case (where AR and CLR are equivalent). We conclude that no case can be made for using 2SLS t-tests. The CLR and ACT tests should be adopted instead, and used in conjunction.

Table 12. ACT Test Rejection Rates and Power with Three Instruments ($\rho = 0.8$) (%)

C	6.90	13.01	40.91	110.55	360.26
$F_{5\%Crit}$	6.93	10.00	22.30	50.00	142.50
$\beta = 0$:					
Total Rejection Rate	0.050	0.050	0.049	0.051	0.051
When $\hat{\beta} > 0$	0.026	0.026	0.025	0.026	0.025
$\beta = -0.3$:					
Total Rejection Rate	0.083	0.150	0.458	0.873	1.000
When $\hat{\beta} < 0$	0.078	0.148	0.458	0.873	1.000
$\beta = 0.3$:					
Total Rejection Rate	0.109	0.172	0.465	0.881	1.000

Note: The table reports the frequency of rejecting the null hypothesis H_0: $\beta = 0$ using the conditional t-test.

14. CONCLUSION

We have examined the behavior of 2SLS given different levels of instrument strength, focusing primarily on the a basic *iid* normal environment. In that context, Staiger-Stock suggested the popular rule of thumb that the first-stage F should be at least 10 for 2SLS to give reliable results. And, in the case of a single instrument, Stock-Yogo showed that a first-stage F of 16.4 ensures maximal size distortion in two-tailed 2SLS t-tests is no more than 5%. However, we find 2SLS is very poorly behaved in environments characterized by first-stage F-statistics in the 10 to 16.4 range.

The problem is not the Stock-Yogo analysis itself, which is perfectly correct, but rather that their focus on maximal size distortions of two-tailed t-tests masks other problems with 2SLS. First, 2SLS has very low power if the first-stage F is in the 10 to 16.4 range. Second, *the 2SLS estimator has the unfortunate property that it tends to generate standard errors that are artificially too low precisely when it generates estimates that are shifted most strongly in the direction of the OLS bias.* Consequently, nearly all significant 2SLS estimates are severely shifted towards OLS when instruments are weak.

In fact, we find standard 2SLS t-tests have little power to detect true negative effects when the OLS bias is positive, even when instruments are "strong" by conventional standards (e.g., F=10). This is of great practical importance, as it means there is little chance of detecting negative program effects given positive selection on unobservables.

A general consequence of the association between 2SLS estimates and their standard errors is that size distortions in one-tailed t-tests are far greater than size distortions in two-tailed tests. We find very high levels of instrument strength are needed to reduce those size distortions to modest levels. For example, if the first-stage F meets the 104.7 threshold suggested by Lee et al. (2020), then 2SLS has reasonable power properties, and size distortions in two-tailed t-tests are modest. But size distortions in one-tailed t-tests are still enormous. In fact, we find a first-stage F-threshold of about 10,000 is needed to eliminate size distortions in one-tailed 2SLS t-tests.

The literature on 2SLS seems to have overlooked the problem of size distortions in one-tailed tests. Applied researchers rarely use one-tailed tests as they expect two-tailed tests to be symmetric (so a two-tailed 5% test is equivalent to a one-tailed 2.5% test). But that is completely false with 2SLS: Even with moderately strong instruments almost all 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 asymmetry in 2SLS t-tests is highly relevant for applied work. Consider the classic problem of estimating the effect of education on wages. The usual concern is that unmeasured ability biases the OLS education coefficient upward. Our results imply that if the OLS bias is indeed positive, then larger positive 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.

We find the Anderson-Rubin (AR) test suffers from the same problem but to a much lesser degree: That is, when instruments are weak the AR statistic tends to be greater when $\hat{\beta}_{2SLS}$ is shifted in the direction of the OLS bias. So AR over-rejects H$_0$: $\beta = 0$ when $\hat{\beta}_{2SLS}$ is shifted towards $E(\hat{\beta}_{OLS})$. Fortunately, however, the problem with the AR test becomes negligible at a much lower first-stage F threshold than for the t-test. Thus, we advise using the AR test even if the first-stage F is in the thousands.

We present an application to estimating "excess sensitivity" of consumption to income using PSID data to assess relative performance of AR and t-tests in a realistic setting. In this example the first-stage F-statistic 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. As Andrews et al. (2019) note, the AR test is robust to weak instrument problems and has (weakly) greater power than any alternative test in just-identified models. Given the weight of the empirical and theoretical evidence, it is clear the AR test should be widely adopted in lieu of the t-test even when instruments are strong.

The asymmetric conditional t-test (ACT) of Mills et al. (2014) corrects the asymmetry in standard t-tests, so we advise it should also be widely adopted in applied work. When the OLS bias is positive the ACT test has greater power to detect true positive effects than the AR test, and *vice versa*, so we advise using both tests in conjunction.

Going beyond the focus on test statistics, we argue that a limitation of most prior work on weak instruments is that the quality of 2SLS estimates is evaluated in isolation, asking how strong instruments ought to be for 2SLS itself to exhibit acceptable statistical properties. In practice applied researchers face a choice between using 2SLS and OLS. So an alternative is to ask "How strong must instruments be for 2SLS to give more reliable results than OLS?" Given commonly used thresholds for testing weak instruments, we find probabilities that 2SLS will perform worse than OLS are substantial. For example, given a first stage F of 10, and given a uniform prior on the degree of the endogeneity problem, we calculate a 52% probability that 2SLS will generate an estimate of β further from the truth than OLS. But if the first-stage F is 50 this figure drops to only 17%.

Given these results, we advise applied researchers to think seriously about reasonable priors on the extent of endogeneity before assessing first-stage F thresholds. We give practical guidance on how to do this. For example, in the classic example of estimating the effect of education on wages, we show that a first-stage F threshold of 50 or better is required to have high confidence that 2SLS will outperform OLS. In cases where such a threshold cannot be met, the use of OLS combined with a serious attempt to control for sources of endogeneity may be a superior research strategy to reliance on IV.

We also evaluate alternatives to 2SLS, including the Fuller and JIVE estimators and the Unbiased estimator of Andrews and Armstrong (2017). If first stage F is 50 then 2SLS, Fuller and Unbiased behave very similarly, while JIVE is inferior. If the first stage F is lower and the level of endogeneity is moderate then none of these estimators is likely to outperform OLS. The Fuller and Unbiased estimators offer significant improvements over 2SLS (and OLS) when endogeneity is severe and instruments are weak.

Finally, we consider the over-identified case with a single endogenous variable. General conclusions carry over from the exactly identified case. In fact, the use of multiple instruments increases the covariance between 2SLS estimates and their standard errors. This makes it more essential to use robust test statistics like AR, ACT and the conditional likelihood ratio (CLR) test in lieu of the t-test, even when instruments are strong.

Interestingly, if the OLS bias is positive, the ACT test has better power to detect a true positive effect than the CLR test, and *vice versa.* As both tests have correct size even when instruments are weak, we conclude they can advantageously be used in conjunction (recalling that CLR and AR are equivalent in the single instrument case). No case can be made for using 2SLS t-tests in either exact or over-identified models.

In conclusion, we note that recent papers by Andrews et al. (2019) and Young (2020) 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 are very grateful to Isaiah Andrews, Peter Phillips and Robert Moffitt for valuable comments, and to Marcelo Moreira for providing us with 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 Mathematical 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. and T. Armstrong (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics 8*(2), 479–503.

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.

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

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. (2004). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'économique 36*(4), 767–808.

Fuller, W. (1977). Some properties of a modification of the limited information estimator. *Econometrica 45*(4), 939–953.

Keane, M. and T. Neal (2021). Weak instrument robust inference for the Frisch labor supply elasticity. *UNSW Economics Working Paper 2021-07b*. Available at SSRN: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3915528`.

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

Kleibergen, F. and R. Paap (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of econometrics 133*(1), 97–126.

Lee, D., J. McCrary, M. Moreira, and J. Porter (2020). Valid t-ratio inference for IV. *arXiv preprint arXiv:2010.05058*.

MaCurdy, T. E. (1981). An empirical model of labor supply in a life-cycle setting. *Journal of political Economy 89*(6), 1059–1085.

Mariger, R. P. and K. Shaw (1993). Unanticipated aggregate disturbances and tests of the life-cycle consumption model using panel data. *The Review of Economics and Statistics*, 48–56.

Marsaglia, G. (2006). Ratios of normal variables. *Journal of Statistical Software 16*(4), 1–10.

Mikusheva, A. (2013). Survey on statistical inferences in weakly-identified instrumental variable models. *Applied Econometrics 29*(1), 117–131.

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. J. Moreira (2019). Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors. *Journal of Econometrics 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.

Olea, J. L. M. and C. Pflueger (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics 31*(3), 358–369.

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.

Phillips, P. C. B. (1983). Exact small sample theory in the simultaneous equations model. *Handbook of Econometrics 1*, 449–516.

Richardson, D. H. (1968). The exact distribution of a structural coefficient estimator. *Journal of the American Statistical Association 63*(324), 1214–1226.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics 2*, 881–935.

Sawa, T. (1969). The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *Journal of the American Statistical association 64*(327), 923–937.

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 Education.

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

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

Young, A. (2020). Consistency without inference: Instrumental variables in practical application. *Working Paper, London School of Economics*.

APPENDICES

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

Consider the following just-identified *iid*-normal linear IV model:

$$\begin{gathered} y_i = \beta x_i + u_i \\ x_i = \pi z_i + e_i \quad \text{where} \quad e_i = \rho u_i + \sqrt{1-\rho^2}\eta_i \\ u_i \sim iidN(0,1), \eta_i \sim iidN(0,1), z_i \sim iidN(0,1) \end{gathered} \tag{A1}$$

The power of both the AR and t-tests depends on three parameters: the true β, the degree of endogeneity ρ, and the population t-statistic on z in the first-stage regression, which we denote λ (= square root of population F). The power of the AR test (i.e., rate of rejecting H_0:β=0 as a function of the true β) is simply:

$$Power_{AR} = \Phi(\lambda D - z_{1-\alpha/2}) + \Phi(-z_{1-\alpha/2} - \lambda D) \tag{A2}$$

where Φ is the standard normal cdf, $D = \beta/\sqrt{1+2\rho\beta+\beta^2}$, and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. We set $\alpha = 0.05$.

To obtain the power function of the of the t-test we follow the analysis in Stock and Yogo (2005), Lee et al. (2020) and Angrist and Kolesár (2021). The power of the two-tailed 2SLS t-test is given by the integral:

$$\begin{aligned} &Power_t = \\ &\int_{-\infty}^{\infty} \left(\mathbb{I}\{t^2 \geq (1-\rho_0^2) z_{1-\alpha/2}^2\} f(t,D,\lambda,\rho_0) + \mathbb{I}\{t^2 \geq z_{1-\alpha/2}^2\} \right) \phi(t-\lambda) dt \end{aligned} \tag{A3}$$

where ϕ is the standard normal density, $\rho_0 = (\rho+\beta)/\sqrt{1+2\rho\beta+\beta^2}$, and:

$$f(t,D,\lambda,\rho_0) = \Phi\left(\frac{a_2 - \lambda D - \rho_0(t-\lambda)}{\sqrt{1-\rho_0^2}}\right) - \Phi\left(\frac{a_1 - \lambda D - \rho_0(t-\lambda)}{\sqrt{1-\rho_0^2}}\right), \tag{A4}$$

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

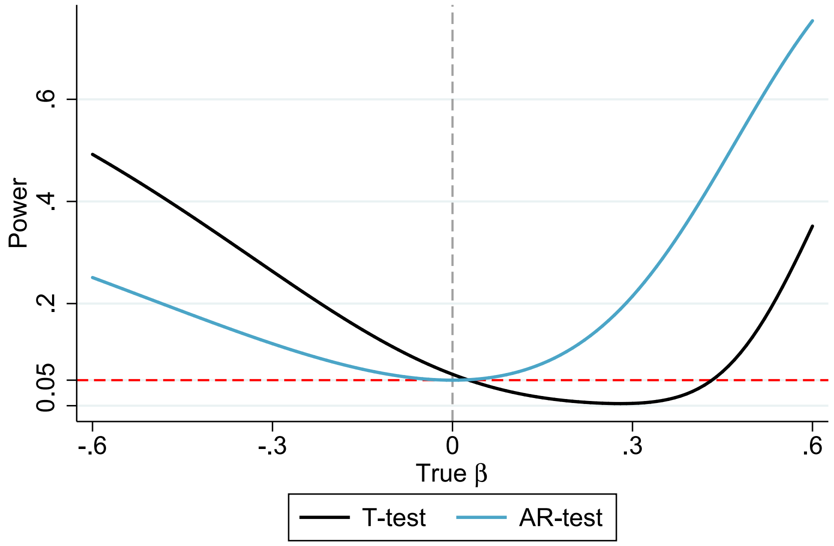
$$a_2 = \frac{\rho_0 z_{1-\alpha/2}^2 t + |t| z_{1-\alpha/2} \sqrt{t^2 - (1-\rho_0^2) z_{1-\alpha/2}^2}}{z_{1-\alpha/2}^2 - t^2}.$$

The integral in (A3) must be evaluated numerically.

A.1. Power of the AR vs t-test at Different Levels of Instrument Strength

We now present power comparisons between the AR and t-test. We take parameter values from an empirical application to estimating the Frisch elasticity in Keane and Neal (2021). The Frisch elasticity is the labor supply response to predictable wage changes. Following the approach developed by MaCurdy (1981), we estimate it via a 2SLS regression of log hours changes on log wage changes, instrumenting for wage growth using an ability test score known as the ASVAB. This is motivated by the fact that higher ability workers have predictably faster wage growth. We estimate $\rho = -0.7$, implying the OLS estimate of the Frisch elasticity is biased in a negative direction. The first-stage F-statistic is 10.12, so we set $\lambda = \sqrt{F} = 3.186$. Results for this case are reported in Figure A1.

Figure A1. Power of the T-Test vs. AR-Test when F = 10.12 ($\rho = -0.7$)



Note: Probability a 5% level test rejects H_0:β=0, conditional on each true β listed on the x-axis. We set $\rho = -0.7$ and the population first-stage F-statistic to 10.12.

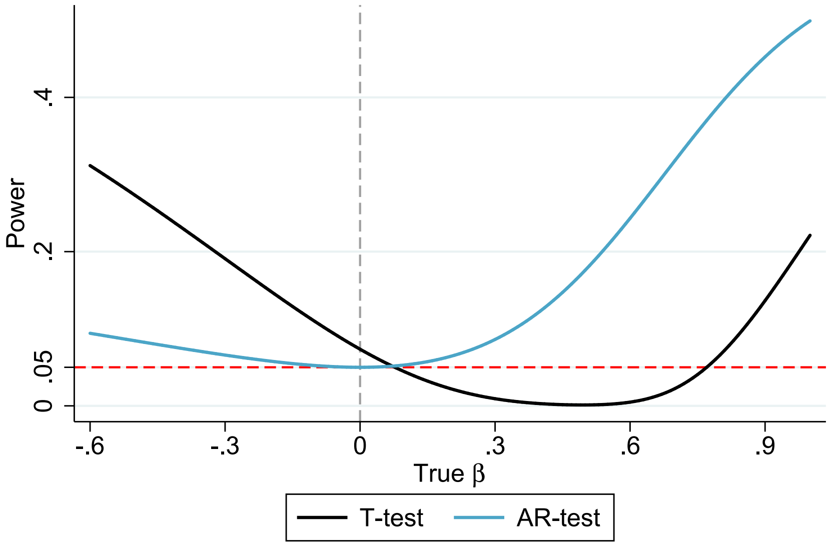
The severe power asymmetry of the t-test is evident in Figure A1. Because the OLS bias is negative ($\rho = -0.70$), the t-test has little power to detect a wide range of true positive β values. In fact, the t-test has power less than size for true effects in the 0.05 to 0.45 range, so it is uninformative over that range. Recall that β in equation (5) is roughly the std. dev. change in y induced by a one std. dev. change in x, so an effect size of 0.45 is substantial in most applications.

Thus, Figure A1 illustrates the t-test's poor power to detect a true β opposite in sign to the OLS bias, even if instrument strength is comfortably above conventional weak IV testing thresholds like $\hat{F} > 10$. A first-stage $\hat{F}$ of at least 23.2 is required to have 95% confidence that population F is at least 10.12.

Figure A1 also illustrates the superior power properties of the AR test. First, the AR test is unbiased: It's power is appropriately minimized when the true β is 0 – in contrast to the t-test whose power is minimized when true β is roughly 0.30. Second, the power of the AR test is far superior when the true β is positive. For example, it reaches about 60% when true elasticity is 0.50, compared to only 15% for the t-test.

Third, the AR test has correct size. It's power evaluated at β=0 is correctly 5%, compared to 6.2% for the the t-test. Another notable feature of Figure A1 is that the t-test appears to have better power than the AR test for negative values of true β. This reveals the flip side of the power asymmetry problem: In samples where the 2SLS estimate is shifted in the direction of the OLS bias, which in this case is negative, the 2SLS standard error is spuriously small, which inflates the power of the t-test. This is not a desirable property, as the standard error exaggerates the precision of the estimate in such cases.

Figure A2. Power of the T-Test vs. AR-Test when $F = 2.3$ ($\rho = -0.7$)



The Staiger-Stock rule of thumb suggests a first-stage $\hat{F}$ of at least 10 indicates an acceptable level of instrument strength. An $\hat{F}$ of 10 gives 95% confidence that population F is at least 2.3. But Figure A2 reveals the power of the 2SLS t-test is very poor in the F=2.3 case. It has essentially no power to detect positive elasticities in the plausible 0.1 to 0.8 range, as power is less than the 5% size of the test throughout this range. The AR test has better power to detect true positive elasticities, but its power is still rather low (e.g., it doesn't pass 20% until the elasticity exceeds 0.5). So while this level of instrument strength is deemed acceptable by common practice, the data is not very informative. We advise that a higher standard of instrument strength be required in practice.

This naturally leads us to ask how large F must be for the t-test to exhibit acceptable power for plausible elasticity values. Figure A3 reports results for a population F of 29.44. A first-stage $\hat{F}$ of at least 50 gives 95% confidence that population F is at least this large. At this level of instrument strength power is much more acceptable: For instance, the power of both tests approaches one when as true elasticity approaches 0.6. However, the t-test still has power less than size for elasticities in the 0.0 to 0.2 range, and very poor power compared to the t-test for elasticities in the 0.0 to 0.4 range.

Finally, Figure A4 considers the case of true F=73.75, a high level of instrument strength. A first-stage $\hat{F}$ of at least 104.7 is required to have 95% confidence that population F is at least this large.[25] Here the power curves of the two tests are much more

[25] We choose to examine this case because Lee et al. (2020) show that a first-stage sample $\hat{F}$ of at least 104.7 is required for the worst-case size distortion in the t-test to be no more than 5%. Their analysis is subtly different from Stock and Yogo (2005), in that their "worst case" refers to the maximum size distortion over all possible values of endogeneity ρ and all possible values of the true F. The worst case scenario for ρ is again ±1, while the worst case for F is $[\hat{F}/(\sqrt{\hat{F}} + 1.96)]^2 = 73.74$.

Figure A3. Power of the T-Test vs. AR-Test when $F = 29.44$ ($\rho = -0.7$)

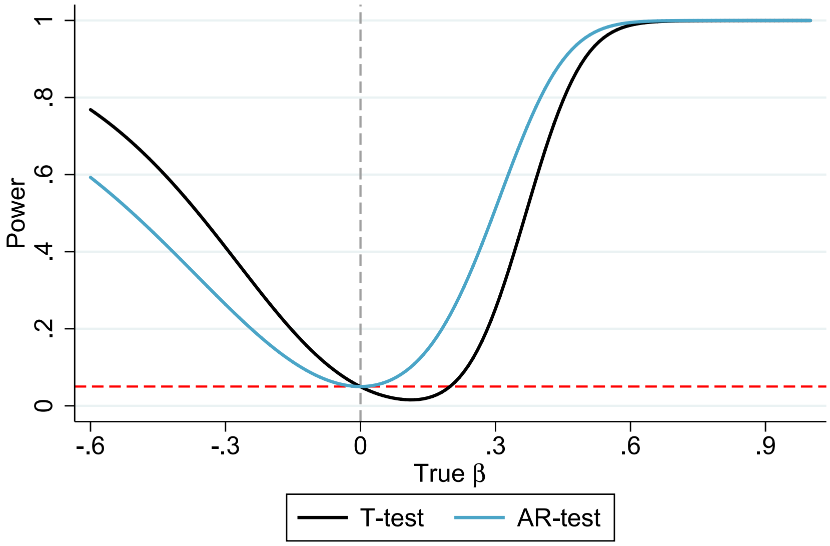
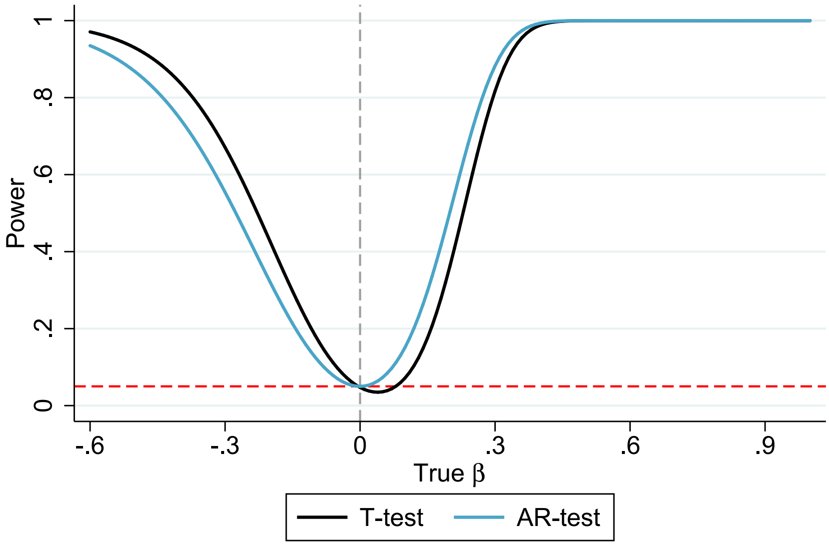


Figure A4. Power of the T-Test vs. AR-Test when $F = 73.75$ ($\rho = -0.7$)



similar, and power of both tests approaches 1 for β around 0.40. But the power advantage of the AR test is still evident in the $\beta \in (0.0, 0.30)$ range. For instance, for β=0.15 the AR test power is 40% vs. 25% for the t-test.

In summary, these results show that the power advantage of the AR test over the t-test can be substantial at empirically relevant elasticity values, even when instruments are quite strong. The power asymmetry of the t-test (i.e., its low power to detect plausible positive elasticities) is dramatic when instruments are weak but persists even when instruments are very strong.

B. SIMULATING THE DISTRIBUTION OF THE T-TEST

Following Mills et al. (2014), if we assume that $\beta_0 = 0$ and define $\Omega = \left(\begin{smallmatrix} \sigma_u^2 & \rho\sigma_u\sigma_e \\ \rho\sigma_u\sigma_e & \sigma_e^2 \end{smallmatrix}\right)$ it is true 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),[26] $t_{1'}$ is the t-statistic of a regression of $x_i - \hat{\rho} y_i$ on z_i, and σ_{2SLS} is the standard error of the 2SLS regression. Furthermore, $c_{11} = \sigma_u$, $c_{12} = 0$, $c_{21} = \rho\sigma_e$, and $c_{22} = \sigma_e\sqrt{1-\rho^2}$. The expression for β_{2SLS} can be further derived as:

$$\beta_{2SLS} = \frac{c_{11}c_{21}t_{AR}^2 + (c_{12}c_{21} + c_{11}c_{22})t_{AR}t_{1'} + c_{12}c_{22}t_{1'}^2}{c_{21}^2 t_{AR}^2 + 2c_{21}c_{22}t_{AR}t_{1'} + c_{22}^2 t_{1'}^2} \tag{B2}$$

Using these equations, it is possible to simulate the distribution of the 2SLS t-test conditional on the strength of the instrument and its correlation with the structural errors using the following procedure:

1 Draw a simulated value of t_{AR} from the $N(0,1)$ distribution.[27]
2 Calculate β_{2SLS} as above but using the simulated value of t_{AR}.
3 Calculate t_{2SLS} using the values from the first and second step.
4 Repeat Steps 1-3 N times.
5 To obtain critical values for a 5% level test, calculate the 2.5% and 97.5% percentile of the N simulated values of t_{2SLS}.

The percentiles from the fifth step form the 5% conditional t-test critical values for each side of the $\beta = 0$ null hypothesis.

[26] 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 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.

[27] In the case of a single instrument, t_{AR} is standard normal. In the case of k instruments it is drawn from $N(0,1) + \sqrt{\chi(k-1)}$.

C. SUPPLEMENTARY TABLES AND FIGURES

Table C1. Median Standard Error for $\hat{\beta}_{2SLS}$

Concentration Parameter ("True First-Stage F")	F critical value to reject C<c at 5%	Standard Error 2SLS
1.82	8.96	0.799
2.30	10.00	0.705
3.84	13.00	0.533
5.78	16.38	0.429
10.00	23.10	0.322
73.75	104.70	0.117

Note: The worst-case OLS bias is 1.0 when $\rho = 1$ and $\pi = 0$.

Figure C1. 2SLS Power Function, t-test of $H_0 : \beta = 0$ when true $\beta = 0.3$

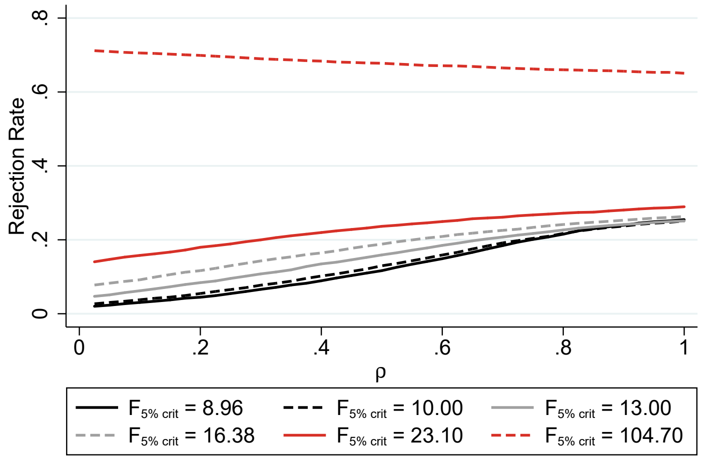


Figure C2. The Standard Error of Optimal IV Plotted Against $\hat{\beta}_{OPT}$ $(\rho = 0.80)$

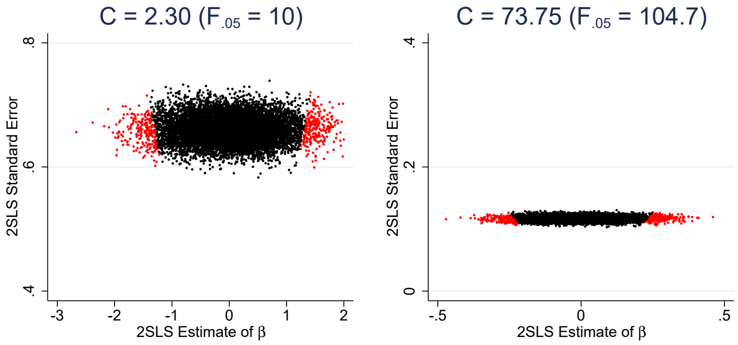
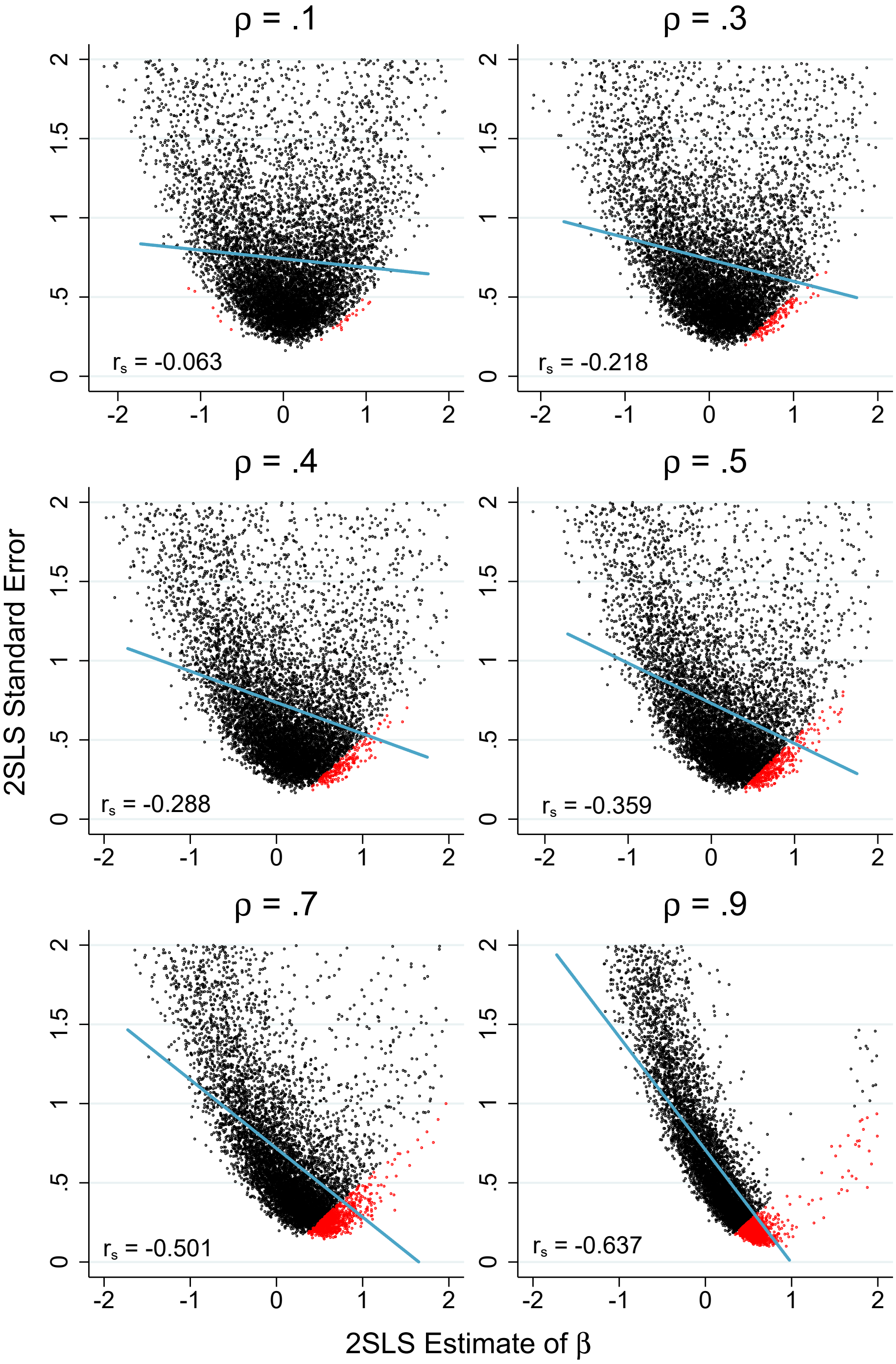
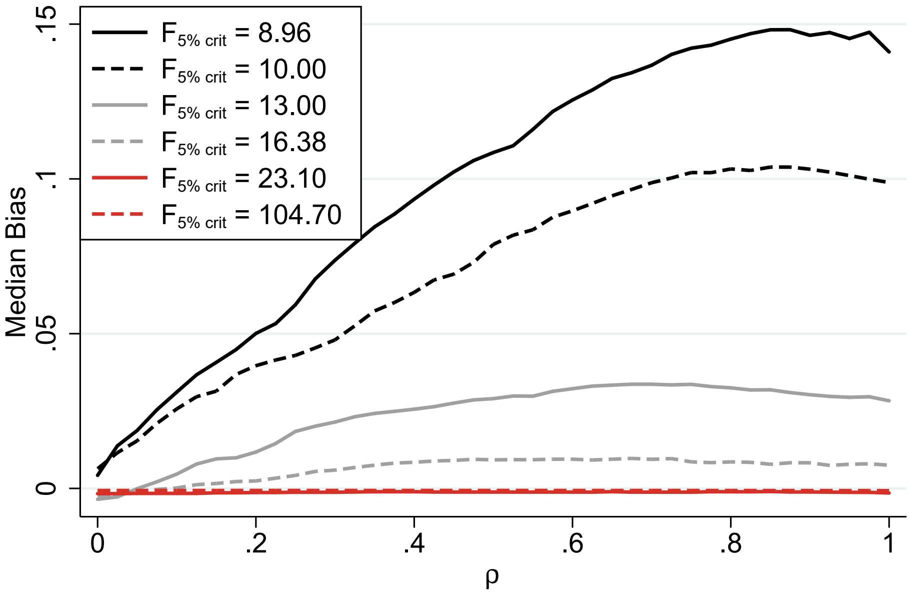


Figure C3. The Covariance of 2SLS Estimates and their Standard Errors



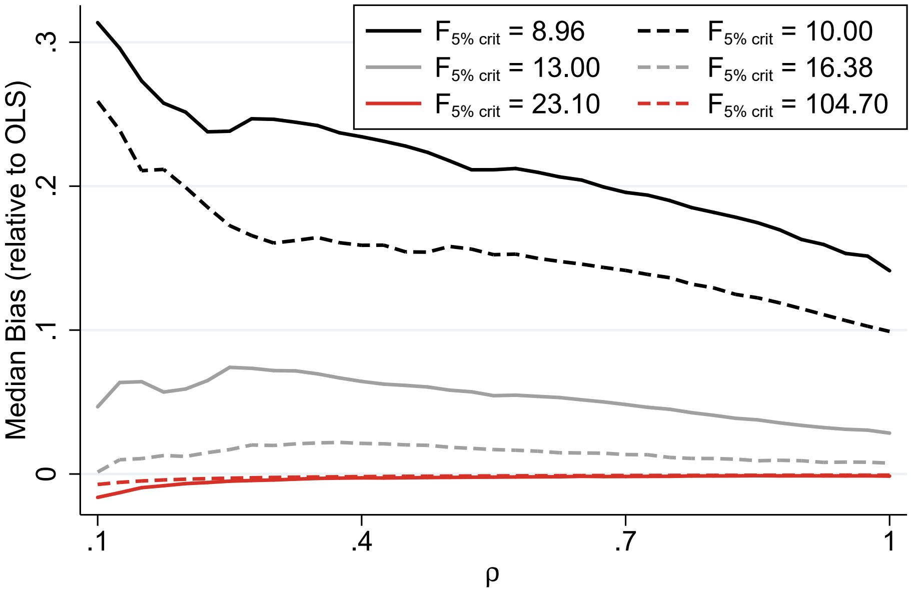
Note: This figure plots the Standard Error of $\hat{\beta}_{2SLS}$ against $\hat{\beta}_{2SLS}$ itself ($C = 2.3, F_{.05} = 10$). Red dots indicate $H_0 : \beta = 0$ rejected at 5% level. The OLS regression line for $SE(\hat{\beta}_{2SLS})$ against $\hat{\beta}_{2SLS}$ is presented in blue and excludes standard errors > 2, while the Spearman correlations r_s do not.

Figure C4. Median Bias of 2SLS by Instrument Strength and ρ



Note: We plot the median of the $\hat{\beta}_{2SLS}$ estimates. The worst-case OLS bias is 1.0 when $\rho = 1$ and $\pi = 0$.

Figure C5. Median Bias of 2SLS Relative to OLS Bias (%)



Note: We plot median($\hat{\beta}_{2SLS}$)/median($\hat{\beta}_{OLS}$). The worst-case OLS bias is 1.0 when $\rho = 1$ and $\pi = 0$.