ARC Centre of Excellence in Population Ageing Research

Working Paper 2021/14

A New Perspective on Weak Instruments

Michael Keane and Timothy Neal

A New Perspective on Weak Instruments

Michael Keane[†] and Timothy Neal[†]

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

Summary It is well-understood that 2SLS has poor properties if instruments are exogenous but weak. We clarify these properties, explain weak instrument tests, and study how behavior of 2SLS depends on instrument strength. A common standard for acceptable instruments is a first-stage F-statistic of at least 10. But we show 2SLS has poor properties in that context: Besides having little power, *2SLS generates artificially low standard errors precisely in those samples where it generates estimates most contaminated by endogeneity.* This problem persists even when instruments are very strong, causing one-tailed 2SLS t-tests to suffer from severe size distortions unless F approaches 10,000. The Anderson-Rubin test alleviates this problem, and should be used even with strong instruments. A first-stage F of 50 or more is necessary to give reasonable confidence that 2SLS will outperform OLS. Otherwise, OLS combined with controls for sources of endogeneity may be a superior research strategy to IV.

JEL: *C12, C26*

Keywords: *Instrumental variables, weak instruments, 2SLS, endogeneity, F-test, size distortions of tests, Anderson-Rubin test, conditional t-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 testing, as well as 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, and to show that these problems persist even in large samples with instruments that would be considered "strong" by conventional standards. This leads us to advocate a higher standard of instrument strength in applications of instrumental variables.

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 tests of the hypothesis $H_0 : \beta = 0$ are likely to reject the null at a rate 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 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 on the case of a single endogenous variable and a single instrument. This case 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 extremely low power, combined with *a very unfortunate tendency to generate artificially low standard errors precisely when it generates estimates most contaminated by endogeneity.*

This covariance between 2SLS standard errors and estimates, 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 covariance between 2SLS standard errors and estimates continues to have 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.

[1] Stock and Yogo (2005) also considered the criterion of bias relative to OLS. But this criterion can only be assessed given over-identification of degree 2. The large majority of applied IV papers use exactly identified models, so the bias criterion is not relevant.

We find that these approaches do alleviate the problems we identify *provided* the instruments are strong enough that 2SLS has acceptable power in the first place, which in practice requires a first-stage F threshold of about 50.

We then examine performance of the main alternative estimators to 2SLS, which are the Fuller (1977) estimator, JIVE and the unbiased estimator of Andrews and Armstrong (2017). We find that Fuller and the unbiased estimator do offer some improvement, but again, it seems their performance cannot be judged adequate unless the first-stage F threshold is at least 50.

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 of 50+ 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.

Finally, we propose a new perspective on weak instruments: Conventional approaches look at 2SLS in isolation, asking how strong instruments must be for it to exhibit acceptable properties. We propose an alternative: Applied researchers face a choice between 2SLS and OLS, so it is natural to ask if 2SLS is likely to deliver superior results? We find the first-stage F must be at least 50 to have confidence in such an outcome. Hence, we suspect that in applied contexts where this threshold cannot be met, the use of OLS combined with a serious attempt to control/proxy for sources of endogeneity is likely to be a superior research strategy to reliance on instrumental variables.

2. SOME BACKGROUND ON THE WEAK INSTRUMENT PROBLEM

To clarify the weak instrument problem, and explain weak instrument tests, we focus on a simple case: A single endogenous variable and a single instrument. This is by far the most common case in practice. We also focus on the case of no 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 $\pi \neq 0$. For our exposition, 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:

$$e_i = \rho u_i + \eta_i \quad \text{where} \quad cov\,(\eta, u) = 0,\ cov\,(z, \eta) = 0 \tag{2.1}$$

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 $\hat{}$ denotes a sample value:

$$\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i x_i} = \beta + \frac{\sum_{i=1}^{n} z_i u_i}{\sum_{i=1}^{n} z_i x_i} = \beta + \frac{\widehat{cov}\,(z, u)}{\widehat{cov}\,(z, x)} \tag{2.2}$$

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 $\pi\sigma_z^2 \neq 0$. So $\hat{\beta}_{2SLS}$ converges to the

true β. But this is not useful in practice: We are interested in properties of 2SLS in **finite** samples – including, as we will see, very large finite samples.

We can substitute (2.1) into (2.2), and write $\hat{\beta}_{2SLS} - \beta$ in the instructive form:

$$\hat{\beta}_{2SLS} - \beta = \frac{\widehat{cov}\,(z,u)}{\pi\widehat{var}\,(z) + \widehat{cov}\,(z,e)} = \frac{\widehat{cov}\,(z,u)}{\pi\widehat{var}\,(z) + \widehat{cov}\,(z,\eta) + \rho\widehat{cov}\,(z,u)} \tag{2.3}$$

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 $\rho = 0$.

In contrast to the simple OLS world, the presence of the $\widehat{cov}\,(z,\eta)$ and $\rho\widehat{cov}\,(z,u)$ terms in the denominator of (2.3) make the finite sample properties of 2SLS very difficult to analyze. In fact, they cause 2SLS to have very odd properties in finite samples.

In this section we explain what these oddities are, why they are serious problems if instruments are "weak," but only academic curiosities when instruments are "strong." We avoid mathematical formalism throughout, in order to give simple intuitive explanations.

We focus exclusively on the case of a perfectly exogenous instrument z with population covariance $cov(z,u) = 0$. We want to concentrate on problems created by weak instruments, abstracting from problems created by departures from instrument exogeneity.

The crucial point to remember is that in finite samples even a perfect instrument has some sample covariance with the error in the structural equation. That means $\widehat{cov}(z,u)$ will be always be non-zero, even if our 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 $\widehat{cov}(z,u)$ and $\widehat{cov}(z,\eta)$ appear in the denominator of (2.3) has unpleasant consequences for the finite sample behavior of the 2SLS estimator - even in large samples. We can learn a lot about these properties of 2SLS just by carefully studying equation (2.3).[2] Now we list some key properties:

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, rendering conventional t-tests unreliable. To see why, assume $\pi > 0$, so z has a positive *population* covariance with x, and $\rho > 0$, so the endogeneity bias is positive. 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 (2.3) 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

[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 (2.3) 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.

amplified (attenuated) when $\widehat{cov}(z,u) < 0$ (>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 $\rho\widehat{cov}(z,u)$ generates both an estimate shifted towards OLS and a low standard error - as it causes a large $\widehat{cov}(z,x)$. *So 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 covariance between 2SLS estimates and 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:

$$\hat{\beta}_{2SLS} - \beta = \frac{\widehat{cov}(z,u)}{\widehat{cov}(z,x)} = \frac{\widehat{cov}(z,u)}{\rho\widehat{cov}(z,u) + \widehat{cov}(z,\eta)} \tag{2.4}$$

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 (2.4) 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 four properties, the bias of the median 2SLS estimate toward OLS has received substantial attention in the applied literature; see 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 (see Phillips 1983, Rothenberg 1984), and Nelson and Startz (1990) and Mikusheva (2013) provide nice expositions. But, we argue, the problems created by covariance between 2SLS estimates and 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 (2.3), so a natural reaction is to assume they will vanish in a sample that is 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, having an instrument that is significant at ***only*** the 5% level in a large sample is not good enough to make the problems with 2SLS vanish. With a "weak" instrument the problematic finite sample properties of 2SLS remain relevant even in very large samples. For 2SLS to have nice properties, large samples alone are not adequate. What we need is an instrument that is "strong," which is 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 the sample correlation is driven by $\pi\widehat{var}(z)$, which reflects a true relationship between z and x, and $\widehat{cov}(z,e)$, 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. The weak instrument test 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.

Finally, recall that $F = NR^2/(1-R^2)$. 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 a normal distribution as $N \to \infty$ is very slow when the first-stage R^2 is small.

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:

$$\begin{gathered} y_i = \beta x_i + u_i \\ x_i = \pi z_i + e_i \quad \text{where} \quad e_i = \rho u_i + \sqrt{1-\rho^2}\eta_i \\ u_i \sim N(0,1), \eta_i \sim N(0,1), z_i \sim N(0,1) \end{gathered} \tag{3.5}$$

Suppose 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 likely to be severely biased towards OLS.

In an important paper, Staiger and Stock (1997) developed the idea that for an instrument to be sufficiently "strong" for 2SLS to have acceptable finite sample properties, it must meet a higher standard than mere 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 sample F-statistic, which have denoted by $\hat{F}$. 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:

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

The 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 so that weak instruments are not a concern?

Formally, Stock and Yogo (2005) derived how C determines the maximum rejection rate (size) of a 2SLS 5% level two-tailed t-test, compared to the correct 5% rejection rate

– where the maximum is over all values of ρ. They argued that this is a natural way to assess how well-behaved the 2SLS estimates are in a finite sample.

For example, suppose you find it acceptable that the 2SLS two-tailed t-test will reject the null H_0: $\beta = 0$ at the 5% level no more than 15% of the time. In other words, you are willing to tolerate a size distortion of 10%. They derive that you need C of 1.82. Suppose you want to bring the size distortion down to just 5% (i.e., your 5% t-test should reject H_0: $\beta = 0$ no more than 10% of the time). Then you need C=5.78. Finally, Lee et al. (2020) ask how big C needs to be so that your 5% level two-tailed t-test is really a 5% level t-test. They derive $C = 73.75$.

Thus, by requiring C to be large enough we can render size distortions of two-tailed 2SLS t-tests as small as desired. But we can't observe C, so we must rely on the sample $\hat{F}$ as an estimate of C. Unfortunately, because C is equal to $Var(z\pi)/\sigma_e^2$ times a factor of N, the 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. This means if we want 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 the 95% level) 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 that level of c.

Table 1 lists several different levels of C, the levels of π required to achieve them when $N = 1000$, and the first-stage F-test critical value for a 5% test that C attains the desired level. For example, as we show in the last row of Table 11, to be confident (at the 95% level) that C is at least 73.75, we need a first-stage $\hat{F}$ of at least 104.7.

Table 1 also shows, in some key cases, the 5% two-sided t-test size achieved by that level of C. For example, Lee et al. (2020) show that if $C = 73.75$ then a 2SLS 5% level two-tailed t-tests achieves the correct 5% rejection rate. Table 1 also includes C=2.3, which corresponds to the Staiger-Stock rule of thumb $\hat{F} > 10$.[5] Note that size distortion depends only on C, not on N and π *per se*, so how we set N and π is somewhat arbitrary.

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 $H_0 : \beta = 0$
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%

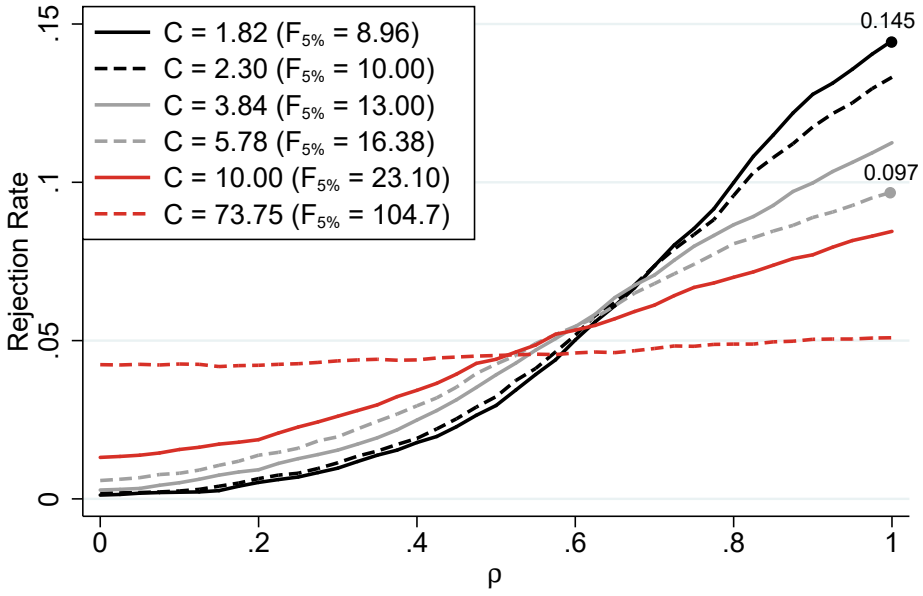
Note: The instrument z is significant at the 5% level in the first stage if F>3.84.

[5]We include $C = 10$ because both Staiger and Stock (1997) and Angrist and Pischke (2008) derive formulas indicating that, when the first moment exists, the bias of 2SLS relative to OLS is roughly $1/[1 + C]$. So $C = 10$ renders the relative bias only 10%. This result is only useful with 3 or more instruments, so the moments of the 2SLS estimator exist. But we felt it was interesting to examine behavior of the 2SLS median in the $C = 10$ case.

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 (3.5), 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: $\beta = 0$ 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 $H_0 : \beta = 0$ using a 5% Two-tailed t-test



Note: The rejection rate should be 5% if 2SLS test statistics are well-behaved.

To understand Figure 1, it is important to note that Stock and Yogo (2005) calculate **worst-case** (defined as 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 ($F_{.05}$=8.96) they predict a worst-case rejection rate of 15%, and our simulations show 14.5%. Similarly, for C=5.78 ($F_{.05}$=16.38) Stock-Yogo predict a worst-case rejection rate of 10%, and our simulations show 9.7%. And when C=73.75 ($F_{.05}$=104.7) the worst-case rejected rate is almost exactly 5%, just as Lee et al. (2020) calculate.

4. PROBLEMS WITH 2SLS HIDDEN BY WEAK INSTRUMENT TESTS

An obvious limitation of the Stock-Yogo analysis is that the worst-case rejection rates tell us little about rejection rates at lower levels of ρ. As we see in Figure 1, rejection rates are increasing in ρ, and this effect is stronger the weaker is the instrument. In fact, the results in Figure 1 raise serious concerns about the behavior of 2SLS, even in cases that easily pass standard tests to rule out weak instruments. To see why, we need to understand **why** rejection rates are increasing in ρ. There are two reasons:

First, if ρ is close to 0 then 2SLS estimates are roughly centered at zero, 2SLS standard errors are very large, and the estimator has very little power. This is why the rejection rate is very small for small values of ρ in Figure 1 (except when $F_{.05}$=104.7).

Second, if $\rho > 0$, then, in samples where $\widehat{cov}(z,u)$ happens to be high, 2SLS will tend to generate both high estimates of β and artificially low standard errors. This is a very unfortunate property, as it means ***the 2SLS estimates appear to be spuriously more precise in samples where they are most shifted upward in the direction of the OLS endogeneity bias***. This is what causes 2SLS to over-reject the null hypothesis H_0: $\beta = 0$ when ρ is large, as we see clearly in Figure 1.

Now we examine these two phenomena in more detail: First, we examine the power of the 2SLS estimator by simulating the probability of rejecting H_0: $\beta = 0$ when it is false.[6] We consider two alternative models with $\beta = 0.30$ 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. The results are reported in Table 2.

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

Concentration Parameter	$F_{5\%Crit}$	$\beta = 0.3$			$\beta = -0.3$		
		$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$
1.82	8.96	1.8	11.7	25.5	1.7	0.1	4.2
2.30	10.00	2.4	13.0	25.1	2.2	0.2	3.2
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: $\beta = 0$.

Focusing on the results for the $F_{.05} = 8.96$ through $F_{.05} = 23.10$ cases, three features are notable: First, the power of 2SLS is very low. Second, power is clearly increasing in ρ if the true β is positive. Third, power is much greater when the true β is positive than when it is negative. For example, in the $F_{.05} = 10$ case widely considered an acceptable threshold for a strong instrument, the probability of rejecting the false null H_0: $\beta = 0$ is only 2.4% when the true β is 0.3 and $\rho = 0$. Power increases with ρ, as the rejection rate increases to 13% when $\rho = 0.5$ and to 25% when ρ=1. Strikingly, if the true β is -0.3

[6] We can also see the low power of the 2SLS estimator by examining its standard error. Table A1 reports the median (over simulation runs) of the asymptotic standard error of $\hat{\beta}_{2SLS}$, for each level of instrument strength. The median standard error is not very sensitive to ρ, so we only report the overall median. To take one example, when $C = 5.78$ ($F_{.05} = 16.38$) the median 2SLS standard error is roughly 0.429. To get a sense of just how large that is, recall that the worst-case endogeneity bias in OLS, which occurs when x is perfectly correlated with the error in the outcome equation (i.e., $\rho = 1$ and $\pi = 0$), is 1.0.

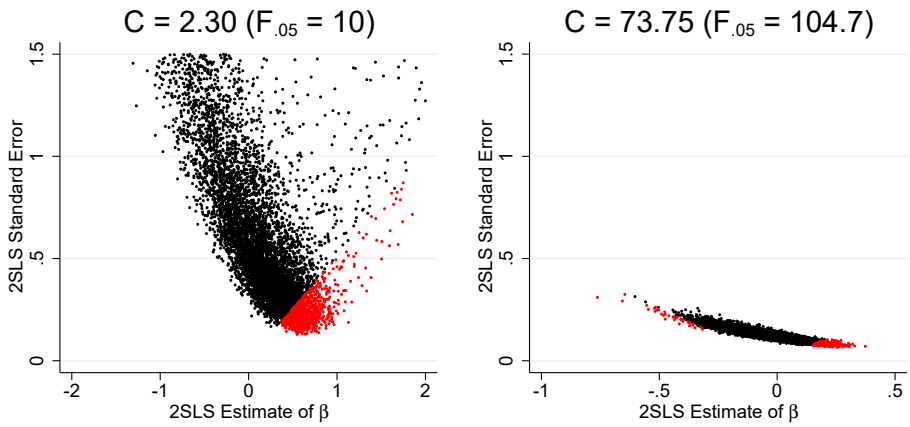
and $\rho = 0.5$ then power is essential zero. This is very concerning, as it means 2SLS has almost no power to detect a (substantial) true negative effect.

The fundamental reason that 2SLS power increases with the degree of endogeneity ρ is the third finite sample property of 2SLS we stressed in Section 2: A positive sample realization of $\rho\widehat{cov}(z,u)$ tends to generate both a 2SLS estimate shifted towards OLS and a low 2SLS standard error. Larger positive realizations of $\rho\widehat{cov}(z,u)$ reduce the 2SLS standard error by increasing $\widehat{cov}(z,x)$. The larger is ρ, the stronger is the tendency for 2SLS standard errors to be artificially small in samples where the 2SLS estimates are most shifted in the direction of the OLS bias.

In the present case $\rho > 0$, so the OLS bias is positive, and hence 2SLS estimates $\hat{\beta}_{2SLS}$ and their standard errors have a negative covariance. This is why the rate that 2SLS rejects H_0 increases with ρ in Figure 1. This pattern is unfortunate, as it means the 2SLS estimates appear more informative precisely when $\hat{\beta}_{2SLS}$ is most shifted in the direction of the OLS bias. The negative covariance between 2SLS estimates and their standard errors also explains why 2SLS has almost no power to detect true negative effects when instrument strength is in the $F_{.05} = 8.96$ through $F_{.05} = 23.10$ range.

Figure 2 shows the covariance between 2SLS estimates and their standard errors is very strong when ρ is large. It plots $se(\hat{\beta}_{2SLS})$ against $\hat{\beta}_{2SLS}$. The left panel shows the case where ρ=0.8 and C=2.30 ($F_{.05}$=10). A strong negative covariance between 2SLS estimates and their standard errors is very evident. As a result, the *2SLS estimates that are most shifted toward the OLS bias appear to be more precise.*

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 covariance 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.[7]

[7]Lee et al. (2020) argue that if $F > 10$ is used as a threshold, the critical value of the t-test can be

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 covariance between $se(\hat{\beta}_{2SLS})$ and $\hat{\beta}_{2SLS}$ persists, but it is weaker. 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. This asymmetry in positive vs. negative rejections is a direct consequence of the negative covariance between 2SLS estimates and standard errors.

Table 3 gives a broader view of this covariance phenomenon by reporting slope coefficients from regressions of $se(\hat{\beta}_{2SLS})$ on $\hat{\beta}_{2SLS}$ for different levels of $C(F_{.05})$ and selected values of ρ. Clearly, the negative relationship between $se(\hat{\beta}_{2SLS})$ and $\hat{\beta}_{2SLS}$ is not specific to the examples in Figure 2. Table 3 shows how the negative relationship gets stronger as ρ increases, which drives the pattern of rejection rates increasing with ρ in Figure 1.

Table 3. Slope Coefficients from Regressions of $se(\beta_{2SLS})$ against $\hat{\beta}_{2SLS}$

Concentration Parameter	F critical value to reject C<c at 5%	Slope when $\rho = 0.2$	Slope when $\rho = 0.5$	Slope when $\rho = 0.8$
1.82	8.9	-0.092	-0.188	-0.441
2.30	10.00	-0.097	-0.254	-0.556
3.84	13.00	-0.126	-0.387	-0.751
5.78	16.38	-0.175	-0.448	-0.772
10.00	23.10	-0.189	-0.480	-0.701
73.75	104.70	-0.052	-0.127	-0.200

Note: Simulations where the standard error exceeding 2 were not included in the calculation.

An obvious alternative explanation for the pattern in Figure 1, which we can rule out, is that bias in the median 2SLS estimate increases as the degree of endogeneity (ρ) increases. In principle, this could cause the rejection rate of H_0: $\beta = 0$ to increase with ρ. However, 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 A3 illustrates this point. If C=10 ($F_{.05}$=23) or better, the 2SLS estimates are essentially median unbiased. Even at C=5.78 ($F_{.05}$=16.4) the median bias is trivial. 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 F of 10 would be quite sufficient.

In summary, Stock-Yogo style weak instrument analysis focuses on the worst-case (over ρ) behavior of two-tailed 2SLS t-tests. This conceals problematic behavior of 2SLS estimators that become apparent if we look over the whole range of ρ, as in Figure 1. 2SLS rejection rates are steeply increasing in ρ (unless $F_{.05}$=105) because 2SLS has very low power when ρ is small, but as ρ increases the 2SLS standard errors become artificially too small in samples where the 2SLS estimates are too high. *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 that are shifted in the direction of the OLS bias* (positive

replaced by 3.43 to achieve a true 5% test. But this correction does not fix the fundamental problem revealed in Figures 1 and 2. In fact, if we use 3.43 as the threshold, the rejection rate is 1.75% rather than 5%, and all rejections occur when $\hat{\beta}_{2SLS}$ is positive.

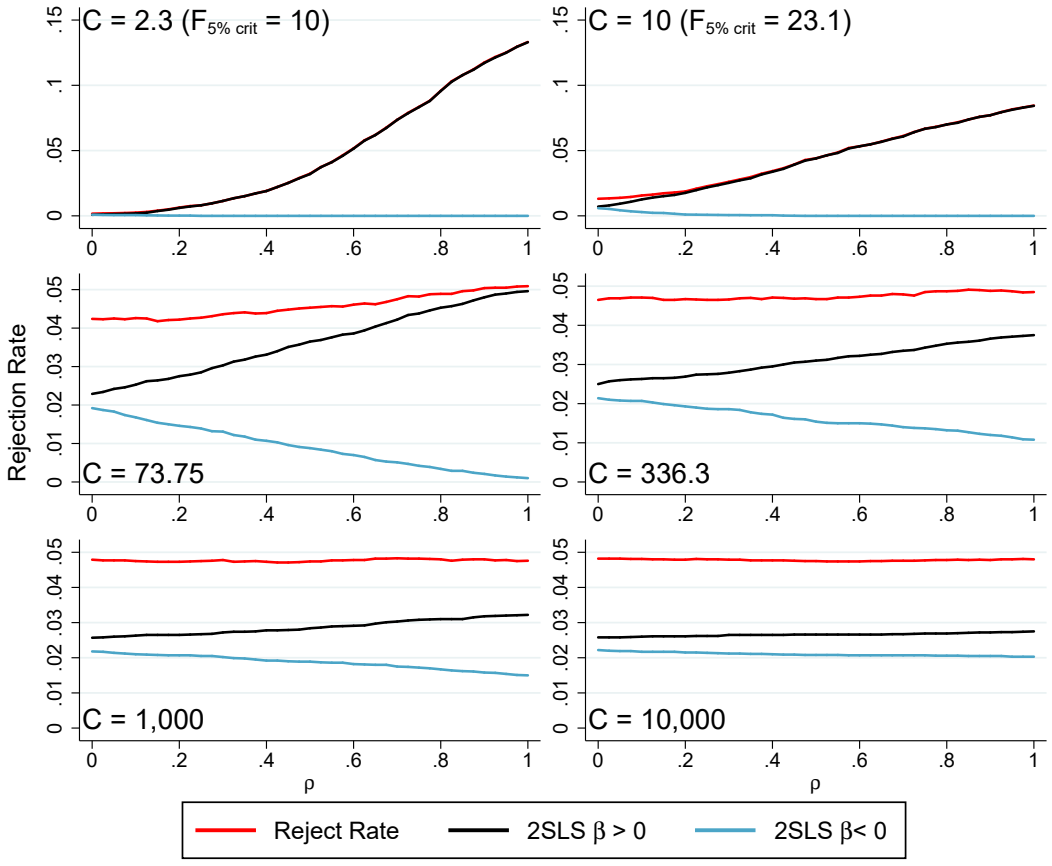
in this case). The fact that 2SLS is approximately median unbiased is not so useful when only estimates shifted in the direction of the OLS bias are likely to be significant, and they are significant far too often.

5. ONE-TAILED TESTS AND ANDERSON-RUBIN TESTS

An important consequence of the covariance between 2SLS estimates and their standard errors is that distortions in one-tailed tests are much greater than distortions in two-tailed tests. This is shown in Figure 3. 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 reject 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 3. Rejection Rates of One and Two-Tailed Tests



In the case of C=74 ($F_{.05}$=105) that Lee et al. (2020) indicate is sufficient to eliminate maximal size distortion, the reject rate of a 5% two-tailed t-test increases only modestly from 4.1% if $\rho = 0$ to 5% when $\rho = 1$. But size distortions in one-tailed t-tests are much greater. The reject 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 reject 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 covariance between $se(\hat{\beta}_{2SLS})$ and $\hat{\beta}_{2SLS}$, which imparts positive $\hat{\beta}_{2SLS}$ estimates with spuriously high precision.

How strong do instruments need to be for rejection rates to be invariant to the level of endogeneity, and for both one and two-tailed tests to have correct size? If the concentration parameter C is increased to 336.3, which corresponds to a first-stage $F_{.05}$=400, the reject rate of two-tailed 5% level t-tests is roughly constant regardless of ρ. But the asymmetry in rejection rates for one-tailed tests is still 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.

A common suggestion in the weak IV literature is to use tests that are robust to weak instruments. In the one endogenous variable, one instrument case the unambiguous suggestion is the Anderson and Rubin (1949) test, which is uniformly most powerful. The AR test is simply the F-test from a regression of y on z. Then, $\hat{\beta}_{2SLS}$ is judged to be significant at the 5% level if that F-test is significant at the 5% level. This test obviously has a "correct" 5% size, as it is simply an F-test from an OLS regression.

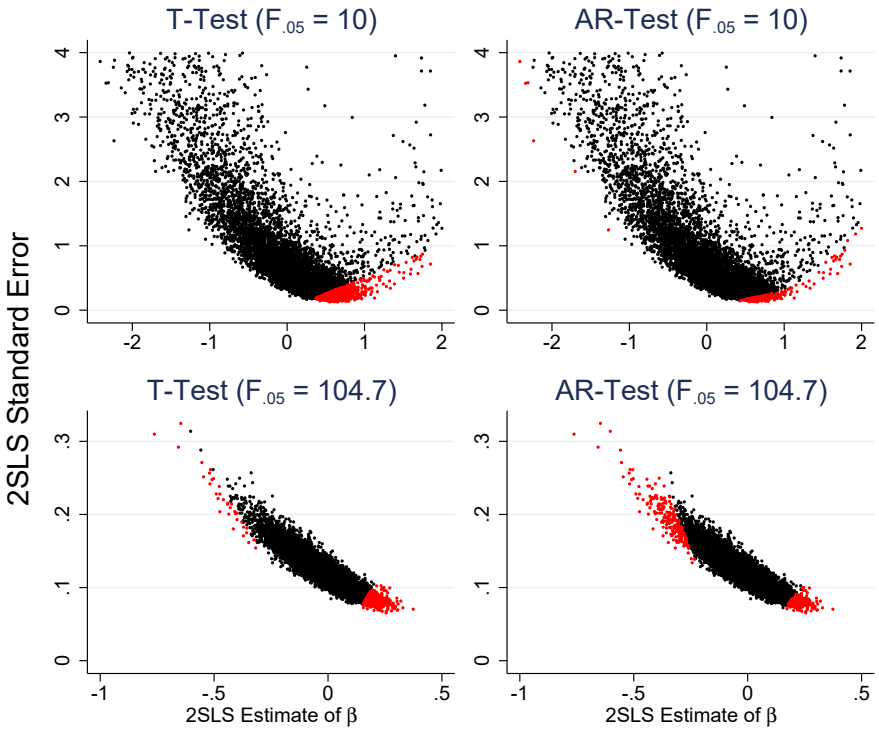
However, if instruments are weak, the AR test is more likely to call 2SLS estimates significant if they are shifted in the direction of the OLS bias. The problem is again the strong positive covariance between $\rho\widehat{cov}(z,u)$ and $\hat{\beta}_{2SLS}$. A large value of $\rho\widehat{cov}(z,u)$ also generates a large value of the AR test. Thus, if $\rho > 0$ the AR test and $\hat{\beta}_{2SLS}$ have a positive covariance. Hence, the AR test is more likely to reject H_0: $\beta = 0$ if $\hat{\beta}_{2SLS} > 0$.

Figure 4 compares results from t-tests vs. AR tests, focusing on the case of ρ=0.80. In the case of $F_{.05}$=10 the 5% two-tailed t-test rejects H_0: $\beta = 0$ at a 10% rate, and all rejections involve $\hat{\beta}_{2SLS} > 0$. The AR test rejects H_0: $\beta = 0$ at a 4.8% rate, which only differs from the correct 5% rate due to sampling variation. However, in the top right of Figure 4, 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.

The size distortion in the two-tailed t-test is mostly eliminated if $F_{.05}$=105. But, as we see in the bottom panel of Figure 4, 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 first-stage F vastly smaller than required for the t-test.

In summary, 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. While both the t-test and AR test cause 2SLS estimates to appear spuriously more precise when they are shifted in the direction of the OLS bias, the AR test is far less sensitive to this problem.

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



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

6. THE CONDITIONAL T-TEST APPROACH

As we have seen, standard 2SLS t-tests give misleading results due to the covariance between 2SLS estimates and standard errors, even when instruments are very strong. This same basic problem affects the AR test, but much less severely. 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 a useful way to address this problem.

Due to the covariance 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.[8]

Table 4 reports summary statistics for critical values simulated using the DGP in (3.5) 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 4 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.

Table 4. 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.

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 covariance 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 5 presents a graphical illustration of how an 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 indicates 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.

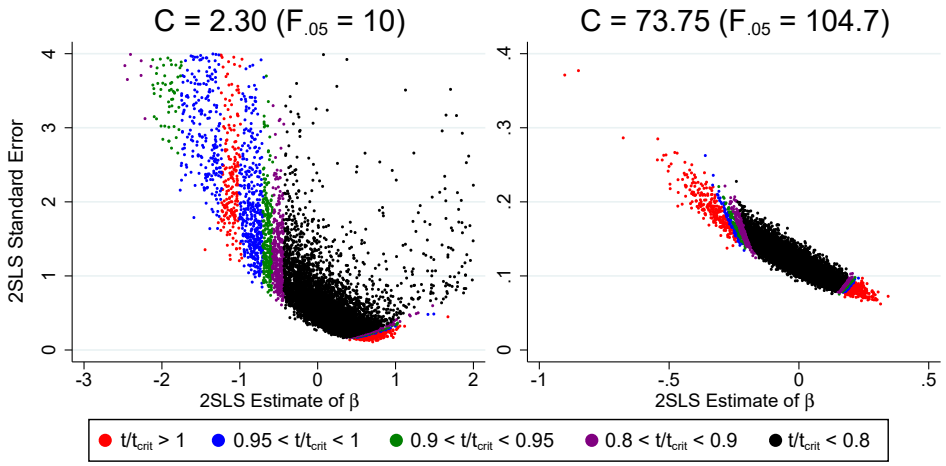
We have observed that standard 2SLS t-tests have little power to detect true negative effects when the OLS bias is positive, even when instruments are "strong" by conventional standards (e.g., F=10). This is of great practical importance, as it means there is little chance of detecting negative program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

[8]Draw many datasets with the same $(\hat{F}, \hat{\rho})$ but with different structural errors, using a construction in Moreira (2003). Then run 2SLS on each dataset, and form the distribution of the t-statistics. Then, for example, use the 95th percentile of the distribution as the conditional critical value for a 5% test. Marcelo Moreira provided us with his Matlab code which does this calculation efficiently.

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).[9] Applied researchers may find it odd to adopt such a "weak" standard, but, given that covariance 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 5 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 5 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.

Figure 5. Mills et al. (2014) One-tailed test rejections ($\rho = 0.80$)



Note: Dots are colored depending on the ratio of the t-statistic with its critical value. Runs with a standard error > 4 are not shown.

In Table 5 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 is example: It is a general 2SLS property that follows directly from the positive covariance between $\rho\widehat{cov}(z,u)$, $\hat{\beta}_{2SLS}$ and the value of the AR test (as we discussed in Section 5).

[9]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.

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 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.[10] But that result only holds within the class of two-tailed tests with symmetric critical values.[11]

Table 5. Rejection Rates for conditional t-tests and AR tests ($\rho = 0.8$) (%)

C	$F_{5\%Crit}$	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: $\beta = 0$.

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 $H_0 : \beta = 0$ 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

[10] 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.

[11] 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 covariance between 2SLS estimates and standard errors that arises in the $\rho > 0$ case.

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.

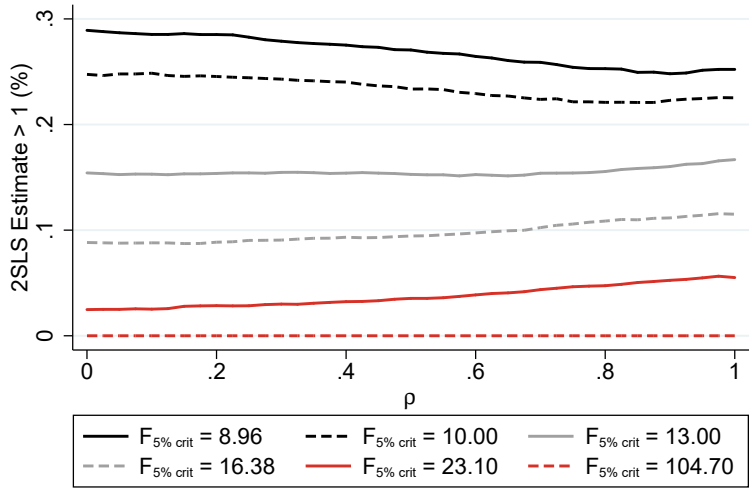
Moving to the case of C=5.780 ($F_{.05}$=16.38) we see some improvement for both tests, as power attains levels of 7.3% to 15%, which clearly exceeds the 5% size level, and "wrong sign" rejections become rare. But these power levels still seem uninspiring. As in the strong instrument case, and for the 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.

7. A NEW PERSPECTIVE ON WEAK INSTRUMENT TESTS

A notable aspect of the weak instrument literature is that performance of 2SLS is typically evaluated in isolation, asking how strong instruments must be for 2SLS to exhibit acceptable statistical properties. But in practice applied researchers face a choice between using 2SLS and OLS. So an alternative way to look at the question is to ask: "How strong do instruments need to be for 2SLS to generate more reliable results than OLS?"

Figure 6. Probability of 2SLS Estimation Error Exceeding Worst-case OLS Bias.



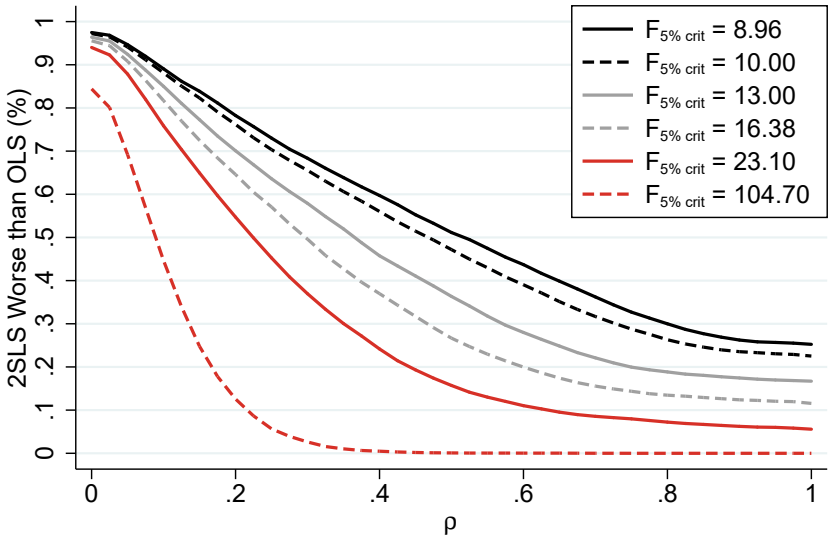
Note: We plot the proportion of times that $|\hat{\beta}_{2SLS}| > 1$, which is the worst-case bias of OLS.

We begin by reporting, in Figure 6, the frequency with which $|\hat{\beta}_{2SLS}| > 1$. To put errors of this size in perspective, recall that 1.0 is the worst-case endogeneity bias in OLS, corresponding to $E(\hat{\beta}_{OLS} - \beta)$ in the case where x is *perfectly* correlated with the error in the outcome equation ($\rho = 1$ and π=0). The results in Figure 6 show that, unless the first stage F is very large, large outlier 2SLS estimation errors of this magnitude are quite common. For example, consider the case where C=3.84 ($F_{.05}$=13). As we saw in

Figure A4, in this case the 2SLS *median* bias is well below 10% of the OLS bias at all levels of ρ. But here we see that there is roughly a 15% chance that $|\hat{\beta}_{2SLS}| > 1$.

Next in Figure 7, 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%.

Figure 7. Probability of 2SLS Performing Worse than OLS



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

The results in Figure 7 are hard to assess without a prior on reasonable values of the parameter ρ, which determines the extent of the endogeneity problem. As we noted earlier, the worst-case for endogeneity bias in OLS is the case where x is *perfectly* correlated with the error in the outcome equation, $\rho = 1$ and π=0. Values of ρ close to one may be relevant in some macro applications (e.g., a regression of consumption on income), but in typical applied micro applications one would never expect ρ to be so high.

For example, consider a regression of log wages on education. It would be surprising to get an R^2 as high as 0.25, meaning the correlation between education and log earnings is certainly no higher than 0.5. 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 0.5.

Table 6 presents the probability that 2SLS will perform worse than OLS, under different scenarios for the concentration parameter C (and the associated first-stage F), and different priors on ρ. For example, in the case of C=2.30 ($F_{.05}$=10), which corresponds to the Staiger-Stock rule of thumb for strong instruments, and given a uniform prior $\rho \in [0, 1]$, the probability that OLS outperforms 2SLS is 52%. Alternatively, we have

argued that in a typical applied micro application - estimating the effect of education on wages – a plausible upper bound on ρ is 0.50. Given a uniform prior $\rho \in [0, 0.5]$, the probability that OLS outperforms 2SLS is 72%. A researcher who thinks education is highly (but not completely) endogenous, might, for example, have a uniform prior of $\rho \in [0.4, 0.5]$. Even in that case, the probability that OLS outperforms 2SLS is 52%.

Table 6. Probability of 2SLS Performing Worse than OLS (%)

Concentration Parameter	$F_{5\%Crit}$	Prior Expectation of ρ: 0 to 1	0 to 0.5	0.4 to 0.5	0.5 to 1
1.82	8.96	55	74	55	35
2.30	10.00	**52**	**72**	**52**	31
3.84	13.00	44	65	41	23
5.78	16.38	38	59	32	16
10.00	23.10	**30**	**50**	20	9
73.75	104.70	11	21	0	0

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.*

Notice that even in the case of C=10.0 ($F_{.05}$=23.1), which is a considerably higher standard of instrument strength than the Staiger-Stock rule of thumb, the probability that OLS outperforms 2SLS is 30% given the uniform prior $\rho \in [0, 1]$, increasing to 50% under the more reasonable prior (for typical applied micro applications) of $\rho \in [0, 0.5]$.

A perhaps surprising result is that even under the very stringent strong instrument requirement of C=73.75 ($F_{.05}$=104.7), a uniform prior on $\rho \in [0, 0.5]$ still generates a 21% risk that 2SLS will give worse results than OLS. Overall, the results of this section suggest that instruments must be much stronger than standard thresholds in the literature - like the popular $\hat{F} > 10$ threshold - would suggest before one can be fairly confident that 2SLS will give results that are superior to OLS, in the sense that $|\hat{\beta}_{2SLS} - \beta| < |\hat{\beta}_{OLS} - \beta|$.

8. HOW BIG DOES F REALLY NEED TO BE?

We find 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 it is likely to underperform OLS, in that $\hat{\beta}_{2SLS}$ is likely to be further from the true value than $\hat{\beta}_{OLS}$. 2SLS still performs poorly if we adopt the more stringent $F_{.05}$=23.1 threshold, but it is very likely to outperform OLS if we adopt a far more stringent $F_{.05}$=105 threshold. This raises the question whether there is some intermediate F-threshold where 2SLS is likely to outperform OLS? We examine this question in Table 7. Based on these results, it appears that raising the F-threshold to roughly 50 attains most of the improvement achievable by going all the way to 105. We caution however, that the AR and ACT tests should be used in lieu of the t-test if F is that small.

Our whole discussion has been centered on the *iid* normal case, in order to focus on key issues.[12] But in assessing acceptable first-stage F statistics in practice it is important to consider the impact of heteroskedasticity. In general, as Andrews et al. (2019) point out,

[12]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).

Table 7. Probability of 2SLS Performing Worse than OLS (%)

Concentration Parameter	$F_{5\%Crit}$	Prior Expectation of ρ: 0 to 1	0 to 0.5	0.4 to 0.5	0.5 to 1
10.00	23.10	30	50	20	9
21.90	40.00	19	36	7	2
29.44	**50.00**	**17**	32	4	1
37.22	60.00	15	29	2	0
73.75	104.70	11	21	0	0

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

there is no theoretical justification for using 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.

9. IS THERE A BETTER ALTERNATIVE TO 2SLS?

We have argued that a first-stage F of at least 50 is necessary to have reasonable confidence that 2SLS will out-perform 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.[13]

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 8.

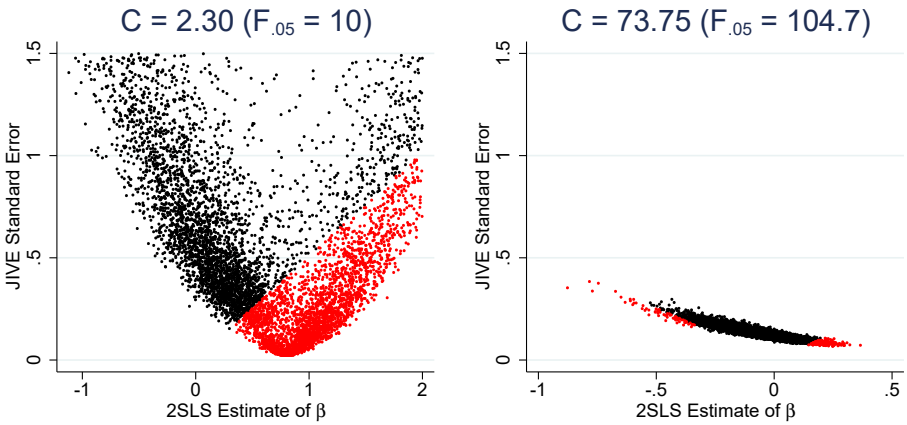
In the case of C=2.30 ($F_{.05}$=10) the JIVE estimator causes us to reject H_0: $\beta = 0$ via a two-tailed 5% t-test a striking 29% of the time, and *all* the rejections are positive. In Sections 4-5 we emphasized the problem that 2SLS is much more likely to judge estimates significant if they are shifted in the direct of the OLS bias. Here we see that JIVE makes this problem much worse. The negative covariance between $se(\hat{\beta}_{JIVE})$ and $\hat{\beta}_{JIVE}$ imparts positive $\hat{\beta}_{JIVE}$ estimates with spuriously high precision.

The reason JIVE performs worse than 2SLS is that 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.[14] In the

[13]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.

[14]In fact, in the runs in the left panel of Figure 8, $\widehat{cov}(z,x)$ 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!

Figure 8. 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.*

right panel of Figure 8 we see that in the relatively strong instrument case of C=73.75 ($F_{.05}$=104.7) 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.

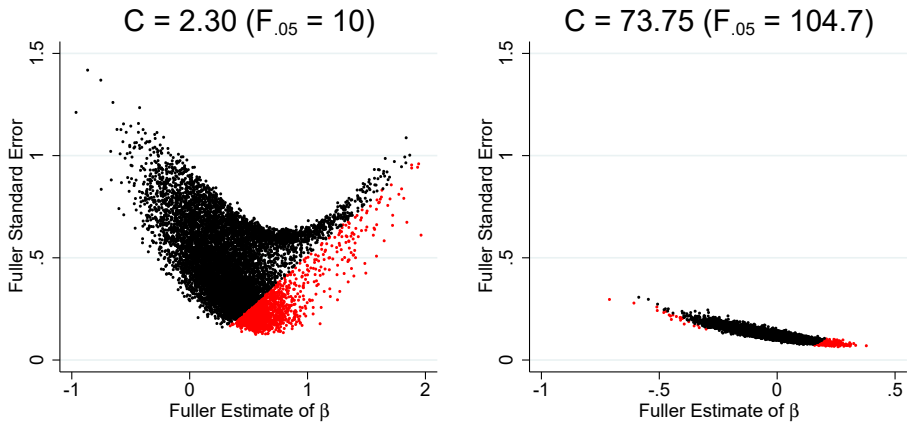
We report results using the Fuller estimator in Figure 9. First, consider the case of $\rho = 0.80$ and C=2.30 ($F_{.05}$=10). Comparison with Figure 2 shows that Fuller estimates and standard errors are substantially less dispersed than 2SLS. Fuller causes us to reject H_0: $\beta = 0$ via a two-tailed 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.

In the right panel Figure 9 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.[15]

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 $1/\pi$.

[15]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.

Figure 9. 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.*

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 $\hat{\beta}_U$.[16]

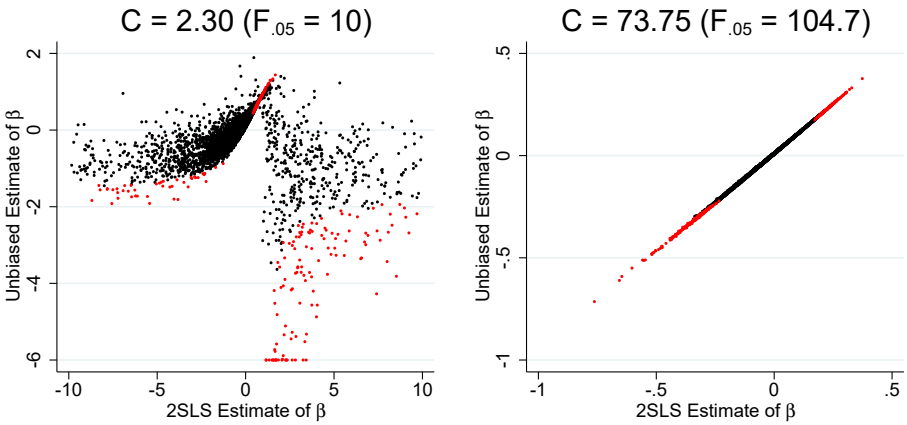
It is important to understand how the first-stage estimate $\hat{\pi}$ is modified by this transformation. Note $\phi/(1-\Phi)$ is the inverse Mills ratio, so $\hat{\pi}^* = E(x|x > \hat{\pi})$ where $x \sim N(0, \sigma_2^2)$. Thus $\hat{\pi}^*$ is positive by construction, and always larger than $\hat{\pi}$. 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 10, 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 10, 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).

[16]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.

Figure 10. 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.

Finally, Table 8 repeats the analysis of Sections 7-8, by asking how often the alternative estimators perform worse than 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 $\rho \in [0, 0.5]$, which we argue is plausible in many applied micro applications, the probability that OLS outperforms 2SLS is 72% when $F_{.05} = 10$. For Fuller the figure is 57% and for Unbiased it is 65%. So while these alternatives outperform 2SLS, their performance can hardly be considered acceptable in any absolute sense when instruments are weak.

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. And all these estimators should be used in conjunction with the AR and ACT tests.

This paper has focused on the single instrument case, but theory shows the performance of 2SLS deteriorates (in that estimates are shifted more towards the OLS bias) if the number of instruments is increased while holding first-stage F fixed. The absolute performance of alternative estimators (LIML, Fuller, JIVE, Unbiased) will also deteriorate, but less so. The implication is that an even higher threshold of instrument relevance (than $F > 50$) would be desirable with multiple instruments.

10. CONCLUSION

We have examined the behavior of 2SLS given different levels of instrument strength, in 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 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%. But we find 2SLS is very poorly behaved in environments characterized by first-stage F-statistics in the 10 to 16.4 range.

Table 8. Probability of Estimators Performing Worse than OLS (%)

Estimator	Prior Expectation of ρ: 0 to 1	0 to 0.5	0.4 to 0.5	0.5 to 1
$C = 2.30$, $F_{5\%Crit} = 10.00$				
2SLS	52	72	52	31
JIVE	68	77	66	59
Fuller	35	57	31	13
Unbiased	40	65	40	16
$C = 10.00$, $F_{5\%Crit} = 23.10$				
2SLS	30	50	20	09
JIVE	43	59	35	26
Fuller	24	45	13	02
Unbiased	25	46	15	04
$C = 29.44$, $F_{5\%Crit} = 50.00$				
2SLS	17	32	04	01
JIVE	19	35	07	03
Fuller	16	31	03	00
Unbiased	16	31	03	00

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 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 severly biased 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 correlation 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 edu-

cation 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: 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. Thus, we advise using the AR test even if the first-stage F is in the thousands.

The conditional t-tests of Mills et al. (2014) correct the asymmetry in standard t-tests, so we advise they should be widely adopted in applied work. When the OLS bias is positive they have much 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. But in practice applied researchers face a choice between using 2SLS and OLS. So an alternative way to look at the issue is to ask "How strong do instruments need to be for 2SLS to generate 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 true value than OLS.

Instead, a first-stage F of 50 or more is necessary to have reasonable confidence 2SLS will outperform OLS. Given these results, we advise applied researchers to adopt a first-stage F threshold of 50 or better, and to rely on AR tests even if F is in the thousands. Given the issues with 2SLS we have discussed, we suspect that in many weak instrument contexts the use of OLS combined with a serious attempt to find controls/proxies for sources of endogeneity may be a superior research strategy to reliance on IV.

Finally, 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. As Young (2020) shows, heteroskedasticity and clustering accentuate the problems we describe.

ACKNOWLEDGEMENTS

We are very grateful to Isaiah Andrews 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

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 J. S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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.

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

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*.

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

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

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

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*.

A. APPENDIX TABLES AND FIGURES

Table A1. 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 A1. 2SLS Power Function, t-test of $H_0 : \beta = 0$ when true $\beta = 0.3$

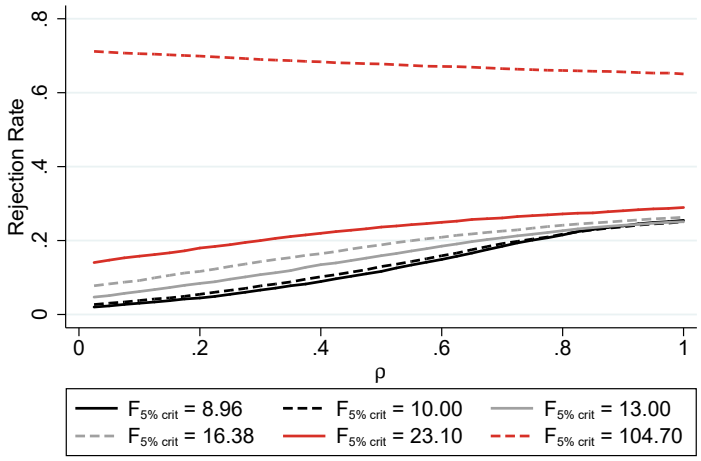


Figure A2. The Standard Error of Optimal IV Plotted Against $\hat{\beta}_{OPT}$ ($\rho = 0.80$)

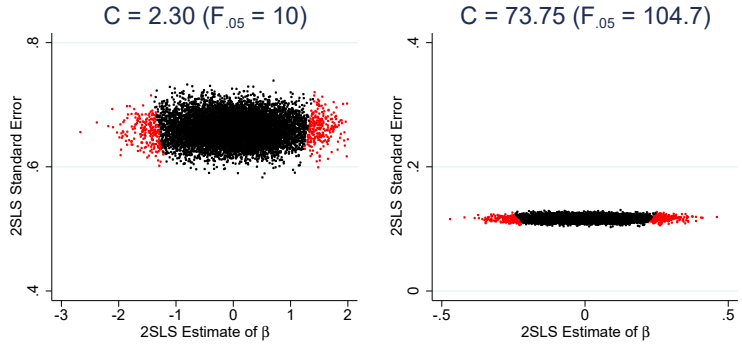
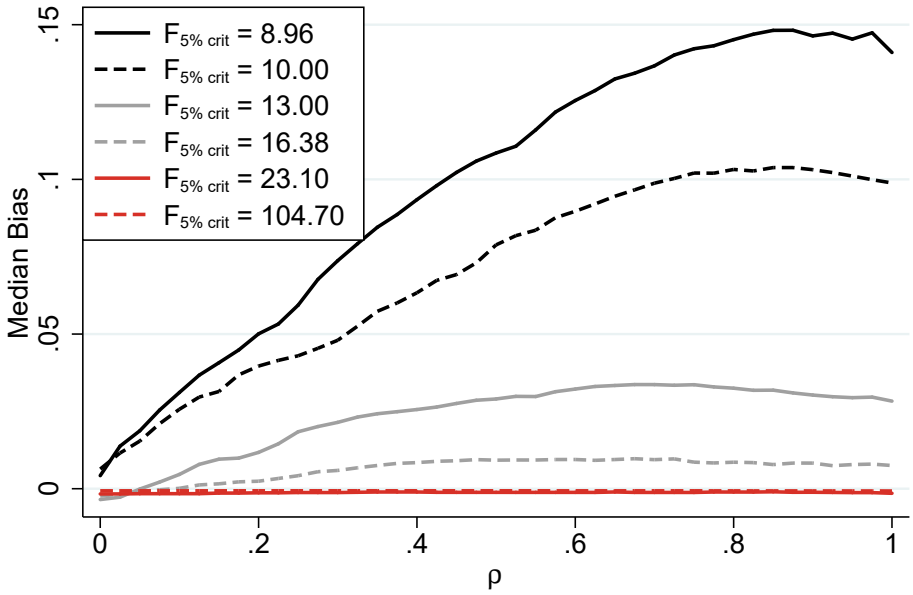
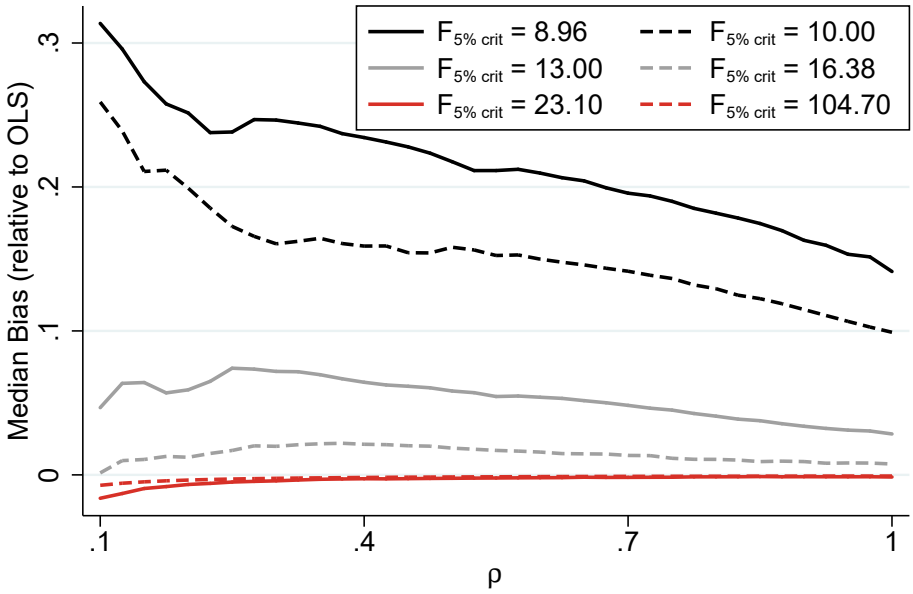


Figure A3. Median Bias of 2SLS by Instrument Strength and ρ



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 A4. 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 $\pi = 0$.