ANNUAL REVIEWS

*Annual Review of Economics*

# A Practical Guide to Weak Instruments

Michael P. Keane[1,2,4] and Timothy Neal[2,3,4]

[1]Carey School of Business and Department of Economics, Johns Hopkins University, Baltimore, Maryland, USA; email: mkeane14@jhu.edu

[2]School of Economics, University of New South Wales, Sydney, New South Wales, Australia; email: timothy.neal@unsw.edu.au

[3]Institute for Climate Risk and Response, University of New South Wales, Sydney, New South Wales, Australia

[4]Australian Research Council (ARC) Centre of Excellence in Population Ageing Research (CEPAR), Sydney, New South Wales, Australia

Annu. Rev. Econ. 2024. 16:185–212

First published as a Review in Advance on April 26, 2024

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-092123-111021



JEL codes: C26, C36, C12, C13



## Keywords

instrumental variables, two-stage least squares, 2SLS, generalized method of moments, GMM, endogeneity, causal inference, size inflation, *t*-test, Anderson-Rubin test, conditional likelihood ratio test, CLR test

## Abstract

We survey the weak instrumental variables (IV) literature with the aim of giving simple advice to applied researchers. This literature focuses heavily on the problem of size inflation in two-stage least squares (2SLS) two-tailed *t*-tests that arises if instruments are weak. A common standard for acceptable instrument strength is a first-stage *F* of 10, which renders this size inflation modest. However, 2SLS suffers from other important problems that exist at much higher levels of instrument strength. In particular, 2SLS standard errors tend to be artificially small in samples where the 2SLS estimate is close to ordinary least squares (OLS). This power asymmetry means the *t*-test has inflated power to detect false positive effects when the OLS bias is positive. The Anderson-Rubin (AR) test avoids this problem and should be used in lieu of the *t*-test even with strong instruments. We illustrate the practical importance of this issue in IV papers published in the *American Economic Review* from 2011 to 2023. Use of the AR test often reverses *t*-test results. In particular, IV estimates that are close to OLS and significant according to the *t*-test are often insignificant according to AR. We also show that for first-stage *F* in the 10–20 range there is a high probability that OLS estimates will be closer to the truth than 2SLS. Hence we advocate a higher standard of instrument strength in applied work.

## 1. INTRODUCTION

The past 35 years have seen an explosion of applied work that uses instrumental variable (IV) methods to deal with endogeneity problems. However, work by Bound et al. (1995) has made applied economists acutely aware that two-stage least squares (2SLS) estimators have poor properties when instruments are exogenous but weak, meaning they are only marginally significant in the first stage of 2SLS. In that case, it is now well understood that 2SLS estimates and standard errors can be very misleading, even in large samples.

Two particularly poor 2SLS properties have received tremendous attention. First, when instruments are weak, the median of the 2SLS estimator is biased toward ordinary least squares (OLS). Second, the 2SLS $t$-test suffers from size inflation, meaning a 5% level $t$-test may reject a true null hypothesis at a rate higher than 5%. In the exactly identified case (one endogenous variable and one instrument), Staiger & Stock (1997) showed that these problems are modest as long as the $F$ statistic for significance of the instrument in the first stage of 2SLS exceeds 10.

In recent work (Keane & Neal 2023) we explained an additional poor property of 2SLS $t$-tests that has not received attention in the prior literature. The $t$-test suffers from a power asymmetry: 2SLS standard errors tend to be artificially small when the 2SLS estimate is close to OLS and large when the 2SLS estimate is far from OLS. The practical consequence is that 2SLS estimates that are near OLS will too often appear significant, even when the true effect is zero. Conversely, the $t$-test has little power to detect true values that are far from OLS. Importantly, this power asymmetry problem remains severe even when the first-stage $F$ is well above levels that are conventionally deemed adequate, that is, $F > 10$.

Fortunately, the power asymmetry problem can be avoided by using the Anderson-Rubin (AR) test (Anderson & Rubin 1949) instead of the $t$-test. In the one instrument case, the AR test of $H_0{:}\beta = 0$ is simply the $t$-test from regression of the outcome $y$ on the predicted value of the endogenous variable $\hat{x}$ obtained from the first stage of 2SLS, where $\hat{x} = \hat{\pi} z$ and $z$ is the instrument.[1] Ironically, this is the naive mistake we teach beginner students of econometrics to avoid, as it does not deliver correct asymptotic standard errors for $\hat{\beta}_{2SLS}$. However, it does deliver the optimal test for significance of $\hat{\beta}_{2SLS}$. The AR test has correct size even when instruments are weak, and it largely avoids the power asymmetry problem.

We illustrate the practical importance of this issue by showing how it affects results from *American Economic Review* (AER) papers published from 2011 to 2023. We found 49 replicable papers where first-stage $F$ was below 50 or not reported, and we found that in 12 of these (24%) the use of AR rather than the $t$-test overturns a key result. In particular, we show there is a systematic pattern whereby 2SLS estimates that are close to OLS are judged significant by the $t$-test but insignificant according to the AR test.[2]

A key point is that the power asymmetry afflicting the $t$-test remains severe even if first-stage $F$ is well above 10 or above the more refined weak IV test thresholds developed by Stock & Yogo (2005). It is not just a weak instrument problem, in the conventional sense of the term. Theorists commonly advocate using AR when instruments are weak (see, e.g., Andrews et al. 2019). The novelty of our message is that we advocate abandoning the $t$-test altogether and using AR even when the instrument is strong. In applications, AR is simple to implement, and it is simple to adapt AR to heteroskedastic and clustered data.

[1] We ignore the presence of other exogenous covariates ($w$) to simplify exposition. Given such $w$, the AR test is simply the $t$-test for the significance of $\hat{x}$ in the regression of $y$ on $\hat{x}$ and $w$.

[2] Four of the 12 papers had multiple instruments, in which case we use the conditional likelihood ratio (CLR) test, which is a natural extension of the AR approach to the overidentified case.

Conventional weak IV tests focus on bias and size inflation; however, what applied researchers really care about is whether 2SLS estimates are more reliable than OLS. In fact, as we show, for first-stage $F$ in the 10–20 range typically deemed acceptable by weak IV tests, the chance that 2SLS will generate an estimate closer to the truth than OLS is remarkably small, unless endogeneity is very severe. Furthermore, the 2SLS $t$-test has such poor power that only estimates near OLS are likely to be significant, due to the power asymmetry. Thus, we argue that more rigorous first-stage $F$ thresholds should be adopted in applied work.

For clarity of exposition we mostly focus on the single instrument case, with only a brief section on multiple instruments. Bound et al. (1995) show that 2SLS bias and $t$-test size problems get worse with multiple instruments. This has led Angrist & Pischke (2008) and Angrist & Kolesár (2024) to advocate using only a single instrument. In other work (Keane & Neal 2023) we show that the power asymmetry problem gets worse with multiple instruments, both for 2SLS and for generalized method of moments (GMM) $t$-tests. However, we find that the conditional likelihood ratio (CLR) test of Moreira (2003), extended to GMM by Kleibergen (2005), can avoid these problems.

The outline of the article is as follows. Section 2 gives a simple explanation of the weak instrument problem, and Section 3 explains weak IV tests. Section 4 explains the power asymmetry problem. Section 5 explains the AR test: We show how it avoids the power asymmetry problem that plagues the $t$-test as well as having other appealing properties. Section 6 discusses the performance of 2SLS relative to OLS. Section 7 presents our analysis of AER papers. Section 8 presents a simple guide for applied researchers, and Section 9 concludes.

## 2. BACKGROUND ON THE WEAK INSTRUMENT PROBLEM

We begin by reviewing how 2SLS works and explaining the various problems created by weak instruments. Assume a researcher is interested in a simple structural equation where the outcome $y$ depends on the single endogenous variable $x$:

$$y = x\beta + u, \quad \text{where} \quad cov(x, u) \neq 0. \tag{1.}$$

For simplicity, we call $\beta$ the effect of $x$ on $y$.[3] The researcher is concerned that a simple OLS regression of $y$ on $x$ will give a biased estimate of $\beta$ because $x$ is endogenous, meaning that $cov(x, u) \neq 0$. A classic example is a regression of wages on education, where the structural error $u$ includes unmeasured ability. If people with higher ability tend to get more education, then $cov(x, u) > 0$, so OLS regression of $y$ on $x$ yields an upward biased estimate of $\beta$.

Fortunately, the researcher also has data on an instrument $z$ with two key properties: It is exogenous, meaning it is uncorrelated with the structural error $u$ in the population, and it is relevant, meaning it is correlated with $x$ in the population. Thus, it is possible to consistently estimate $\beta$ in Equation 1 using an IV estimator.[4] We focus on the 2SLS procedure: In the first stage, we regress $x$ on $z$ to obtain the fitted values $\hat{x}$, and in the second stage we regress $y$ on $\hat{x}$ to consistently estimate $\beta$.

[3]In most cases there are multiple conceptually distinct effects of $x$ on $y$. For example, if $y$ is quantity and $x$ is price, then $\beta$ may represent the effect of price on demand or supply. Similarly, if $y$ is consumption and $x$ is income, then $\beta$ might represent the effect of permanent or transitory income changes. Which effect is identified by an IV estimator depends on the choice of instrument and often on other aspects of the specification as well.

[4]Which effect of $x$ on $y$ is estimated depends on the instrument. For example, to identify $\beta$ in a demand curve, $z$ must only shift the supply curve. To identify effects of permanent income on consumption, $z$ must induce highly persistent income differences across individuals.

We write the first stage of 2SLS, also known as the reduced form for $x$, as

$$x = z\pi + e, \quad \text{where} \quad cov(z, u) = 0,\ cov(z, e) = 0,\ \text{and}\ \pi \neq 0. \tag{2.}$$

The instrument $z$ is exogenous if $cov(z, u) = 0$ and relevant if $\pi \neq 0$. The parameter $\pi$ is most easily thought of as the effect of $z$ on $x$, but mere correlation is adequate for 2SLS.

Equation 2 decomposes $x$ into two parts: The exogenous part $z\pi$ is uncorrelated with the structural error $u$, and the endogenous part $e$ is correlated with $u$. Let $\rho \equiv corr(e, u)$ denote the correlation between the structural and reduced-form errors. The magnitude of $\rho$ determines the severity of the endogeneity problem. $x$ is exogenous if and only if $\rho = 0$.

We can substitute Equation 2 into Equation 1 to obtain what is known as the reduced form for $y$:

$$y = (z\pi)\beta + (\beta e + u). \tag{3.}$$

Assume for a moment we know $\pi$. We may then obtain an unbiased estimate of $\beta$ via an OLS regression of $y$ on $z\pi$. This works because the regressor $z\pi$ is exogenous in Equation 3. That is, $z$ is uncorrelated with both $e$ and $u$, and multiplying $z$ by the constant $\pi$ does not change that. We call the regression of $y$ on $z\pi$ the "infeasible IV" (IIV) estimator.

Of course we do not know $\pi$, so the idea of 2SLS is to replace $\pi$ with the first-stage estimate $\hat{\pi}$ obtained by applying OLS to Equation 2. Then, in the second stage of 2SLS, we regress $y$ on $x\hat{\pi}$ to obtain $\hat{\beta}_{2SLS}$. The second-stage equation can be written as

$$y = (z\hat{\pi})\beta + w, \quad \text{where} \quad w = z(\pi - \hat{\pi})\beta + \beta e + u. \tag{4.}$$

OLS estimation of Equation 4 does not give an unbiased estimate of $\beta$, due to problems created by substituting the data-driven function $\hat{\pi}$ for the constant $\pi$. We explain this in detail in **Supplemental Appendix A**. However, as sample size grows large we have $\hat{\pi} \to \pi$, and hence 2SLS gives a consistent estimator of $\beta$.

Supplemental Material >

An alternative, but numerically equivalent, way to obtain the 2SLS estimator is to (*a*) run an OLS regression of $x$ on $z$ to obtain $\hat{\pi}$; (*b*) run an OLS regression of $y$ on $z$, which, as we see in Equation 3, gives an unbiased estimate $\hat{\pi\beta}$ of $\pi\beta$; and then (*c*) estimate $\beta$ as the ratio $\hat{\beta}_{2SLS} = \hat{\pi\beta}/\hat{\pi}$. Given a sample of observations $\{y_i, x_i, z_i\}$, $i = 1, \ldots, n$ on the random variables $\{y, x, z\}$, the 2SLS estimator of $\beta$ takes the following form:[5]

$$\hat{\beta}_{2SLS} = \frac{\hat{\pi\beta}}{\hat{\pi}} = \frac{\frac{1}{N}\sum_{i=1}^{N} z_i y_i}{\frac{1}{N}\sum_{i=1}^{N} z_i x_i} = \beta + \frac{\frac{1}{N}\sum_{i=1}^{N} z_i u_i}{\frac{1}{N}\sum_{i=1}^{N} z_i x_i} = \beta + \frac{\hat{cov}(z, u)}{\hat{cov}(z, x)}. \tag{5.}$$

Clearly, 2SLS is consistent: As $N \to \infty$ the sample covariance $\hat{cov}(z, u)$ converges to its true value $cov(z, u) = 0$, and $\hat{cov}(z, x)$ converges to $\pi\sigma_z^2 > 0$. So $\hat{\beta}_{2SLS}$ converges to the true $\beta$.

However, the fact that 2SLS is consistent reveals little about its properties and behavior in finite samples—including, as we will see, very large finite samples. We now discuss the properties of 2SLS in finite samples in more detail.

A first point, obvious from Equation 5, is that, in finite samples, the 2SLS estimate $\hat{\beta}_{2SLS}$ departs from the true $\beta$ solely due to finite sample covariance between the instrument and the structural error, $\hat{cov}(z, u) \neq 0$. Although we have that $cov(z, u) = 0$ in the population, it will never be zero in a sample. As we will see, finite sample covariance between the instrument and the structural error is the root cause of several problematic finite sample properties of 2SLS.

[5] In the second stage of 2SLS, the usual OLS formula gives $\hat{\beta}_{2SLS} = \beta + \hat{cov}(z\hat{\pi}, u)/\hat{var}(z\hat{\pi})$. To see that this is equivalent to the expression in Equation 5, note that $\hat{cov}(z, x) = \hat{\pi}\hat{var}(z)$ because $x = z\hat{\pi} + v$ where $\hat{cov}(z, v) = 0$. Then we can write $\hat{cov}(z, u)/\hat{cov}(z, x) = \hat{\pi}\hat{cov}(z, u)/\hat{\pi}\hat{cov}(z, x) = \hat{cov}(z\hat{\pi}, u)/\hat{var}(z\hat{\pi})$.

To explain, it is useful to decompose the first-stage reduced-form error $e$ into parts that are correlated and uncorrelated with the structural error $u$:

$$e = \rho_0 u + \eta, \text{ where } cov(\eta, u) = 0, \ cov(z, \eta) = 0. \quad 6.$$

Here, $\rho_0 \equiv \rho\sigma_e/\sigma_u$ controls the severity of the endogeneity problem. The covariance between the instrument and endogenous variable in a finite sample is

$$\hat{cov}(z, x) = \pi \hat{var}(z) + \hat{cov}(z, \eta) + \rho_0 \hat{cov}(z, u). \quad 7.$$

Thus, $\hat{cov}(z, x)$ has three parts: "good" covariance due to exogenous variation in $x$ generated by $z$; "bad" covariance due to sample correlation of $z$ and $u$, the endogenous part of $x$; and "accidental" covariance due to sample correlation of $z$ and $\eta$, an exogenous part of $x$.

In order to explain some of the problematic finite sample properties of 2SLS, it can be useful to substitute Equation 7 into Equation 5 and write $\hat{\beta}_{2SLS} - \beta$ in the following instructive form:

$$\hat{\beta}_{2SLS} - \beta = \frac{\hat{cov}(z, u)}{\pi \hat{var}(z) + \hat{cov}(z, e)} = \frac{\hat{cov}(z, u)}{\pi \hat{var}(z) + \hat{cov}(z, \eta) + \rho_0 \hat{cov}(z, u)}. \quad 8.$$

Note that the analogous expression for OLS is $\hat{\beta}_{OLS} - \beta = \hat{cov}(x, u)/\hat{var}(x)$. Thus, under standard assumptions, the bias in OLS is $\rho_0 var(u)/var(x)$.

## 2.1. Four Problematic Finite Sample Properties

Given this background, we now list four problematic finite sample properties of 2SLS, along with a simple explanation for each. We focus on the single instrument case.

**2.1.1. Problem 1.** The mean and variance of the 2SLS estimator do not exist, so we cannot define bias. This is simply because it is possible to have a sample realization of $\hat{cov}(z, x) = \pi \hat{var}(z) + \hat{cov}(z, e) \approx 0$, sending the denominator of Equation 8 to zero and causing $\hat{\beta}_{2SLS}$ to explode. Thus, we will focus instead on the median bias of 2SLS.

**2.1.2. Problem 2.** The median of $\hat{\beta}_{2SLS}$ is biased in the direction of OLS if the instrument is weak. To see this, it is useful to first consider the case where the instrument is strong, meaning we can be very confident that $\hat{cov}(z, x)$ is the same sign as $cov(z, x)$ in any sample. Then, from Equation 5, we see that the sign of the sample realization $\hat{cov}(z, u)$ completely determines whether $\hat{\beta}_{2SLS}$ lies above or below the true $\beta$. Of course we expect the random variable $\hat{cov}(z, u)$ to be positive in half of the samples, so the 2SLS estimator is median unbiased.

In contrast, if the instrument is weak, $\hat{cov}(z, x)$ may take the wrong sign in some samples, which induces median bias. To see this, consider the case of $\rho_0 > 0$, so the OLS bias is positive, and assume $cov(z, x) > 0$.[6] Then, from Equation 7, we see that a large negative realization of $\hat{cov}(z, u)$ can drive $\hat{cov}(z, x)$ negative. In that case both the numerator and denominator of Equation 5 are negative, and hence $\hat{\beta}_{2SLS} > \beta$. Thus, we get $\hat{\beta}_{2SLS} > \beta$ in more than half of samples, and the median of $\hat{\beta}_{2SLS}$ is biased in the direction of OLS (positive).

**2.1.3. Problem 3.** The distribution of $\hat{\beta}_{2SLS}$ exhibits skewness and fat tails. To see why, assume again that $cov(z, x) > 0$ and the instrument is strong enough that we are confident that $\hat{cov}(z, x) > 0$, so the denominator of Equation 8 is positive. Take the case $\rho_0 > 0$, so the OLS bias is positive. Then, a negative realization of $\hat{cov}(z, u)$ has two effects: (*a*) It generates a $\hat{\beta}_{2SLS} - \beta < 0$, and

[6] This is without loss of generality, as we can always normalize $z$ so that $\pi > 0$ and hence $cov(z, x) > 0$.

(*b*) it shrinks the denominator of Equation 8, causing that negative estimate to be inflated.[7] The reverse happens if $\hat{cov}(z,u) > 0$. Therefore, estimates that lie above (below) the true $\beta$ are reined in (inflated), and the distribution of $\hat{\beta}_{2SLS} - \beta$ is skewed to the left.[8]

**2.1.4. Problem 4 (power asymmetry).** 2SLS generates artificially low standard errors in samples where $\hat{\beta}_{2SLS}$ is shifted toward $\hat{\beta}_{OLS}$, and artificially high standard errors in samples where $\hat{\beta}_{2SLS}$ is shifted away from $\hat{\beta}_{OLS}$. This was first noted by Keane & Neal (2023).

To understand this problem, consider the 2SLS standard error formula:

$$se(\hat{\beta}_{2SLS}) = \frac{\hat{\sigma}_{2SLS}}{\sqrt{TSS_{x,z}}} \text{ where } \hat{\sigma}_{2SLS} = \sqrt{(N-2)^{-1}\sum_i (y_i - x_i\hat{\beta}_{2SLS})^2}, \quad 9.$$

$$TSS_{x,z} = N \cdot R^2_{x,z} \cdot \hat{var}(x) = N\hat{cov}(z,x)^2/\hat{var}(z) = N\hat{\pi}^2\hat{var}(z). \quad 10.$$

Thus, the 2SLS standard error $se(\hat{\beta}_{2SLS})$ depends on two components: $\hat{\sigma}_{2SLS}$, the standard error of regression, and $TSS_{x,z}$, the total variance of $x$ explained by the instrument $z$. The 2SLS standard error tends to be smaller when $\hat{\beta}_{2SLS}$ is close to $\hat{\beta}_{OLS}$ for two reasons: First, the standard error of regression $\hat{\sigma}_{2SLS}$ is a quadratic function of $\hat{\beta}_{2SLS}$ that is minimized at $\hat{\beta}_{OLS}$. This is simply because OLS minimizes the sum of squared residuals in a regression of $y$ on $x$. Therefore any force that shifts $\hat{\beta}_{2SLS}$ toward $\hat{\beta}_{OLS}$ will reduce the standard error of the 2SLS regression. This is true even in large samples.

Second, the term $1/\sqrt{TSS_{x,z}}$ also tends to be smaller when $\hat{\beta}_{2SLS}$ is close to $\hat{\beta}_{OLS}$. Without loss of generality, assume $cov(z,x) > 0$, and to make the argument transparent, assume the instrument is strong enough that we can be very confident that $\hat{cov}(z,x) > 0$. Then it follows that

1. As we see in Equation 7, a positive sample realization of $\rho_0\hat{cov}(z,u)$ drives up $\hat{cov}(z,x)$ and hence $TSS_{x,z}$, which in turn drives down the 2SLS standard error.
2. As we see in Equation 8, a positive sample realization of $\rho_0\hat{cov}(z,u)$ shifts $\hat{\beta}_{2SLS}$ in the direction of the OLS bias, determined by the sign of $\rho_0$.

Thus, in samples where the instrument is spuriously highly correlated with the endogenous variable, due to sample covariance of $z$ with $u$, the 2SLS estimate is shifted toward OLS and the 2SLS standard error is spuriously small. In fact, as we discuss below, when $\hat{\beta}_{2SLS}$ is close to $\hat{\beta}_{OLS}$, its standard error is actually less than the IIV estimator. By the same logic as above, when $\hat{\beta}_{2SLS}$ is far from $\hat{\beta}_{OLS}$, its standard error is inflated.

Thus, 2SLS has the unfortunate property that it generates artificially low (high) standard errors in samples where $\hat{\beta}_{2SLS}$ is most shifted toward (away from) OLS. This association between 2SLS estimates and their standard errors has important consequences for statistical inference. It means that the 2SLS *t*-test has artificially inflated power to judge estimates to be significant when they shift in the direction of the OLS bias. Conversely, the *t*-test has poor power to detect true negative effects when the OLS bias is positive.

## 2.2. Does a Large Sample Size Solve These Problems?

A large sample size does not solve the power asymmetry problem, as $\hat{\sigma}_{2SLS}$ remains a quadratic function of $\hat{\beta}_{2SLS}$ even in large samples. However, as explained above, the other problems with

[7] A large negative $\hat{cov}(z,u)$ may drive $\hat{cov}(z,x)$ to near zero, generating a large negative outlier.

[8] Things get worse if we allow for the possibility of sufficiently large negative realizations of $\hat{cov}(z,u)$ that $\hat{cov}(z,x)$ is driven negative, as this generates large positive outliers as well.

2SLS are caused by finite sample correlation of the instrument $z$ with the first-stage error $e = \rho_0 u + \eta$.

The sample covariance between the instrument $z$ and the endogenous variable $x$ partly reflects their true relationship $z\pi$, but it is contaminated by spurious correlation that arises because $\hat{cov}(z, e) \neq 0$ in finite samples (see Equation 7). Thus, it is natural to assume these problems will vanish in samples large enough that $\hat{cov}(z, e) \approx 0$.

The error in this logic is that, as sample size grows larger, the value of $\pi$ that is likely to render $z$ significant at the 5% level in the first stage of 2SLS gets small exactly as fast as $\hat{cov}(z, e)$ gets small. As a result, if $z$ is only significant at the 5% level (and not better), $\hat{cov}(z, e)$ remains nonnegligible relative to $\pi \hat{var}(z)$ regardless of sample size. Thus, a large sample alone will not solve the problem. This is a key insight of the weak IV literature [see Staiger & Stock (1997), who develop weak IV asymptotics where $\pi$ shrinks at a $\sqrt{N}$ rate].

## 2.3. What Is a Weak Versus a Strong Instrument? Why Does It Depend on *F*?

At this point we can give intuitive definitions of weak and strong instruments. An instrument is weak if $\pi\, var(z)$ is small enough that $\hat{cov}(z, e)$ remains nonnegligible relative to $\pi \hat{var}(z)$ even in large samples. Conversely, an instrument is strong if $\pi\, var(z)$ is large enough that we are confident that $|\pi \hat{var}(z)| \gg |\hat{cov}(z, e)|$. In other words, we are confident that the sample correlation between $x$ and $z$ mostly reflects their true relationship and not a spurious correlation arising because $\hat{corr}(z, e) \neq 0$ in finite samples.[9] It is simple to see this is equivalent to requiring the first-stage $F$ to be large in some sense.

The "true" first-stage $F$ that we could construct if we observed $\pi$, $var(z)$, and $\sigma_e^2$ in Equation 2 is defined as

$$F = N\frac{var(z\pi)}{\sigma_e^2} = N\frac{\pi^2\sigma_z^2}{\sigma_e^2}. \qquad 11.$$

Note that $|\pi \hat{var}(z)| \gg |\hat{cov}(z, e)|$ can be rewritten as

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

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

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

Finally, substituting the true values for $\hat{\sigma}_z$ and $\hat{\sigma}_z$, and squaring both sides, we obtain

$$\sqrt{N}\frac{|\pi|\sigma_z}{\sigma_e} \gg k \;\; \rightarrow \;\; F = N\frac{\pi^2\sigma_z^2}{\sigma_e^2} \gg k^2.$$

Thus, our intuitive notion of wanting confidence that $|\pi \hat{var}(z)| \gg |\hat{cov}(z, e)|$ corresponds to a desire to have a first-stage $F$-statistic that is "big" in some sense. As $F = NR^2/(1 - R^2)$, a key insight is that the properties of 2SLS do not depend on $N$ or first-stage $R^2$ per se, but only on how they combine to form $F$. The weak IV testing literature asks just how big the first-stage $F$ needs to be for 2SLS to have nice properties. We explain these tests next.

[9] If $|\pi \hat{var}(z)| \gg |\hat{cov}(z, e)|$, the $\pi \hat{var}(z)$ term in the denominator of Equation 8 dominates the $\hat{cov}(z, e)$ term. Then Equation 8 reduces to just $\hat{\beta}_{2SLS} - \beta \approx \hat{cov}(z, u)/\pi \hat{var}(z)$, 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 $\beta$. So 2SLS is approximately unbiased, and normality is a decent approximation to its sampling distribution.

Table 1 First-stage $F$ values required to achieve different objectives

| True $F$ | $\hat{F}$ critical value | Max $t$-test size |
|---|---|---|
| 1.82 | 8.96 | 15% |
| 2.30 | 10.00 | 13.5% |
| 5.78 | 16.38 | 10% |
| 10.00 | 23.10 | 8.6% |
| 29.44 | 50.00 | 6.4% |
| 73.75 | 104.70 | 5% |

## 3. A SIMPLE GUIDE TO WEAK INSTRUMENT TESTS

If the instrument is weak, the distribution of the 2SLS estimator is nonnormal, which renders the $t$-test unreliable. One problem that has received a great deal of attention is $t$-test size inflation: If the instrument is weak and endogeneity is severe, a 5% level two-tailed $t$-test of a true null hypothesis $H_0{:}\beta = 0$ will reject at a rate higher than 5%. In an influential paper, Stock & Yogo (2005) derived levels of first-stage sample $\hat{F}$ that are sufficiently high to give confidence that two-tailed $t$-test size inflation is modest.

To develop these sample $\hat{F}$ thresholds, Stock & Yogo utilized a formula for power of the $t$-test in terms of true $F$, $\rho$, and true $\beta$. We explain this formula in detail in other work (Keane & Neal 2023, appendix A). Using this formula to evaluate power at $\beta = 0$ gives the size of the test (i.e., the probability of rejecting $H_0{:}\beta = 0$). Size depends on $\rho$, so Stock & Yogo focus on the maximal size distortion, which occurs when $\rho = \pm 1$—that is, when endogeneity is extremely severe. We present some key calculations in **Table 1**.

For example, if the true $F$ is 1.82, then a 2SLS two-tailed 5% level $t$-test will reject the true null $H_0{:}\beta = 0$ at the inflated rate of 15% in the worst-case scenario where endogeneity is extremely severe. So the maximal size distortion of the test is 10%. Suppose instead you want a maximal size distortion of just 5% (i.e., your 5% $t$-test rejects $H_0{:}\beta = 0$ no more than 10% of the time); then you need $F = 5.78$. By requiring true $F$ to be large enough, one can render the maximal size distortion as small as desired.

Of course, we cannot observe the true $F$ but only the sample $\hat{F}$ statistic for the significance of $z$ in the first stage, defined as $\hat{F} = N\hat{\pi}^2\hat{\sigma}_z^2/\hat{\sigma}_e^2$. Therefore, the weak instrument testing literature uses sample $\hat{F}$ to make inferences about true $F$. Unfortunately, because $F$ equals $var(z\pi)/\sigma_e^2$ times a factor of $N$, the sample $\hat{F}$ is not a very accurate estimate of true $F$, and it does not get more accurate as sample size increases. In particular, regardless of sample size, sample $\hat{F}$ is a draw from a noncentral $F$ distribution with noncentrality parameter equal to the true $F$. For example, to be confident (at the 95% level) that $F$ is at least 1.82, we need $\hat{F}$ to be at least 8.96. Henceforth we write $F_{5\%} = 8.96$.

Staiger & Stock (1997) proposed a popular rule of thumb that first-stage $\hat{F}$ should be at least 10 to be confident the 2SLS estimator is well behaved.[10] This advice has been widely adopted in practice and presented in textbooks. For example, Stock & Watson (2015, p. 490) write: "One simple rule of thumb is that you do need not to worry about weak instruments if the first stage $F$-statistic exceeds 10." As we see in **Table 1**, a first stage $\hat{F}$ of 10 gives 95% confidence that $F$ is at least 2.3, and this implies a maximal two-tailed $t$-test size of 13.5%.

Next, to better understand how weak instrument tests work in practice, we implement a simple simulation experiment. Consider a model with a single endogenous variable $x$ and a single

[10]Note that in the single instrument case this corresponds to a $t$ of 3.16 ($p = 0.0008$).

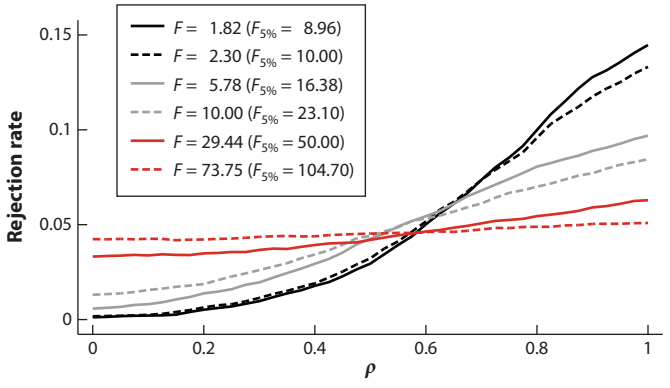


**Figure 1**

Rejection rate of $H_0$:$\beta = 0$ using a 5% two-tailed 2SLS $t$-test, for different values of $F$ and $\rho$. Abbreviations: 2SLS, two-stage least squares.

exogenous instrument $z$. We focus on this simple case as it clarifies the key ideas, and because the single instrument case is very common in applied practice. We have

$$\begin{aligned} y_i &= \beta x_i + u_i, \\ x_i &= \pi z_i + e_i \text{ where } e_i = \rho u_i + \sqrt{1-\rho^2}\eta_i, \qquad i = 1, \ldots, N, \\ u_i &\sim \text{iid}\, N(0,1), \quad \eta_i \sim \text{iid}\, N(0,1), \quad z_i \sim \text{iid}\, N(0,1). \end{aligned} \tag{12.}$$

In this model $\rho \in [-1, 1]$ controls the degree of endogeneity, while $\pi$ determines the strength of the instrument.[11] We normalize $\sigma_z = \sigma_e = 1$ so that $F = N\pi^2$ and the OLS bias is $\rho/(1 + \pi^2) \approx \rho$ for small $\pi$. We generate artificial data sets of size $N = 1{,}000$ and consider six different levels of $\pi$ that generate the six levels of instrument strength that are shown in **Table 1**.[12] We set true $\beta = 0$ and let the degree of endogeneity $\rho$ vary in small increments from 0 to 1.

For each level of $F$ and $\rho$ we simulate 10,000 data sets from the model in Equation 12 and apply 2SLS to the simulated data. We summarize the results in **Figure 1**, which shows the rate at which a two-tail 5% level $t$-test rejects the true null $H_0$:$\beta = 0$ in each case.

The most striking aspect of **Figure 1** is how the rejection rate is strongly increasing as the degree of endogeneity $\rho$ increases. The 2SLS $t$-test is not pivotal, as its size depends on the nuisance parameters $\rho$ and $F$. As we noted earlier, Stock & Yogo (2005) calculate worst-case (maximal) rejection rates over all values of $\rho$. As **Figure 1** shows, the worst case corresponds to $\rho$ near $\pm 1$, so the endogeneity problem is very severe (the figure is symmetric for $\rho < 0$). The results in **Figure 1** agree closely with Stock & Yogo's analysis. For example, for $F = 5.78$ they predict a worst-case rejection rate of 10%, and we obtain 9.7%.

However, **Figure 1** shows that a focus on $\rho = \pm 1$ is not innocuous, as rejection rates vary substantially with $\rho$. In the next section we show how the power asymmetry that we discussed in Section 2 generates this pattern and explain why it is a serious limitation of the $t$-test.

[11] Exogenous covariates may be partialed out of $y$ and $x$ without changing anything of substance. The variance normalizations are without loss of generality as one can standardize $y$, $x$, and $z$.

[12] These are $\pi = 0.0427$, $0.0480$, $0.0760$, $0.1000$, $0.1716$, and $0.2716$—for example, $F = N\pi^2 = 1{,}000(0.10)^2 = 10$.

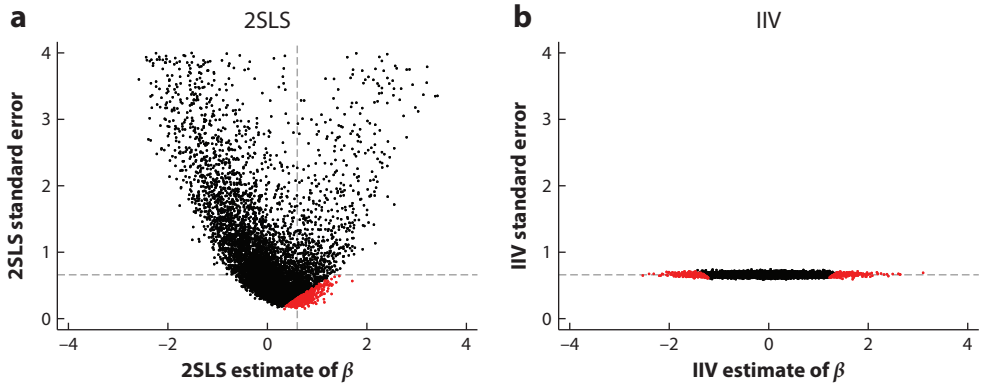


**Figure 2**

Standard error plotted against $\hat{\beta}$ for (*a*) two-stage least squares (2SLS) and (*b*) infeasible instrumental variables (IIV) with $\pi$ known. In each case, $F = 2.3$, $\beta = 0$, and the ordinary least squares (OLS) bias is $E(\hat{\beta}_{\text{OLS}}) = 0.6$. The red dots indicate $H_0$:$\beta = 0$ rejected at 5% level. Runs with standard error >4 are not shown.

## 4. THE POWER ASYMMETRY OF THE 2SLS *t*-TEST

As we explained in Section 2, 2SLS standard errors tend to be smaller when the estimate is shifted toward the OLS bias. **Figure 2** illustrates this power asymmetry phenomenon. We consider the case of $F = 2.3$ ($F_{5\%} = 10$), which corresponds to Staiger & Stock's rule of thumb for acceptable instrument strength, and set $\beta = 0$ and $\rho = 0.6$, so the OLS bias is 0.6. We then simulate 10,000 data sets from this data-generating process (DGP) and apply 2SLS and IIV to each.

**Figure 2*a*** plots the 2SLS standard error $se(\hat{\beta}_{2SLS})$ against the 2SLS estimate $\hat{\beta}_{2SLS}$ for each data set. The strong association between 2SLS estimates and their standard errors is apparent: 2SLS estimates that are close to OLS appear to be much more precisely estimated. However, this precision is spurious; it arises because sample covariance of the instrument with the structural error is relatively high in these data sets, making the instrument appear spuriously strong. As we explained in Section 2, when $cov(z, x)$ is small—it is only 0.048 in this DGP—even a small $\hat{cov}(z, u)$ has big effects!

As a benchmark, **Figure 2*b*** reports results for the IIV estimator we could construct if we knew $\pi$. This is simply an OLS regression of $y$ on $z\pi$. In contrast to 2SLS, an OLS regression generates no association between the estimates and their standard errors (see **Supplemental Appendix B**). In fact, across all data sets the IIV standard errors are tightly clustered around $0.659 = \sigma_u/(N\pi^2\sigma_z^2)^{1/2} = 1/\sqrt{1{,}000} \cdot (0.048)$, the true standard error of the IIV estimator. The empirical standard deviation of the IIV estimates across the 10,000 runs is 0.666, so the IIV standard errors are an accurate reflection of uncertainty.

Supplemental Material >

Strikingly, **Figure 2*a*** shows that when the 2SLS estimate is near OLS, the 2SLS standard error tends to be far below 0.659. In fact, the median 2SLS standard error when $\hat{\beta}_{2SLS}$ is close to $E(\hat{\beta}_{\text{OLS}}) = 0.6$ is only 0.38! It is impossible to gain precision by ignoring information about $\pi$, but this is what 2SLS seems to do. The 2SLS standard error is spuriously small when $\hat{\beta}_{2SLS} \approx 0.6$ due to sample correlation of the instrument with the structural error.

In **Figure 2*a*** we shade in red cases in which the 2SLS estimate is significant according to a two-tailed 5% level *t*-test. A total of 4.9% of estimates are significant, so there is no size inflation. However, all significant estimates are near $E(\hat{\beta}_{\text{OLS}})$, due to the power asymmetry problem. The median significant estimate is 0.70, with a median standard error of 0.29. For IIV, in contrast, significant estimates are equally distributed to the left and right of $\beta = 0$, as expected.

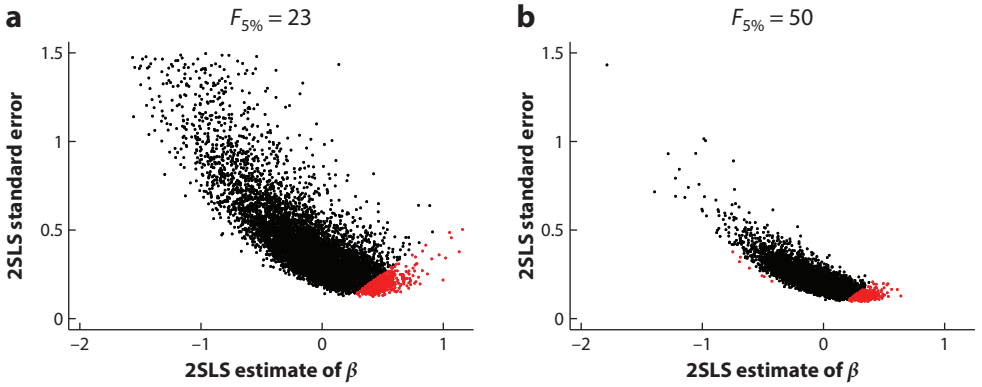


**Figure 3**

Standard error of $\hat{\beta}_{2SLS}$ plotted against $\hat{\beta}_{2SLS}$ with (*a*) $F = 10$ ($F_{5\%} = 23$) and (*b*) $F = 29.4$ ($F_{5\%} = 50$). In both panels, true $\beta = 0$ and the OLS bias is $\beta_{OLS} = 0.6$. Red dots indicate $H_0{:}\beta = 0$ rejected at 5% level. Runs with standard error >1.5 are not shown. Abbreviations: 2SLS, two-stage least squares; OLS, ordinary least squares.

Next, we show that the 2SLS power asymmetry persists at much higher levels of instrument strength. **Figure 3*a*** shows the case of $F = 10$ ($F_{5\%} = 23$), again setting $\beta = 0$ and $\rho = 0.6$. The association between 2SLS estimates and their standard errors is still quantitatively important: In fact, the Spearman's $r_s$ is $-0.576$ and Kendall's $\tau$ is $-0.511$.

The red dots in the figure again indicate cases in which $\hat{\beta}_{2SLS}$ differs significantly from zero according to a two-tailed 5% *t*-test. A total of 5.3% of estimates are significant, so again the size distortion is minor. However, due to the negative association between the 2SLS estimates and their standard errors, all rejections occur when $\hat{\beta}_{2SLS} > 0$, and none occurs when $\hat{\beta}_{2SLS} < 0$. Only the estimates most shifted toward the OLS bias are ever judged significant.

**Figure 3*b*** shows the case of $\rho = 0.6$ and $F = 29.4$ ($F_{5\%} = 50$). Even at this high level of instrument strength, which is far above conventional weak IV testing levels, the negative association between $se(\hat{\beta}_{2SLS})$ and $\hat{\beta}_{2SLS}$ persists. In fact, Spearman's $r_s$ is $-0.92$ and Kendall's $\tau$ is $-0.75$, showing that the power asymmetry is not just a weak instrument phenomenon. The two-tailed 2SLS *t*-test again rejects at close to the correct 5% rate (4.6%), but 98% of those rejections occur when $\hat{\beta}_{2SLS} > 0$. This asymmetry in positive versus negative rejections is a direct consequence of the power asymmetry. The bottom line is that only 2SLS estimates shifted strongly toward OLS are ever likely to be judged significant.

Thus, even with strong instruments, almost all estimates judged significant by two-tailed *t*-tests are shifted toward OLS and not symmetrically distributed around the true value. A key implication, explored in detail in **Supplemental Appendix E**, is that size distortions in one-tailed *t*-tests are much greater than in two-tailed tests. For example, in **Figure 3*a*** a one-tailed 2.5% test of $H_0{:}\beta \leq 0$ rejects at a 5.3% rate. Applied researchers rarely use one-tailed tests, as they expect two-tailed tests to be symmetric (so that a two-tailed 5% test is equivalent to a one-tailed 2.5% test). However, that is completely false with 2SLS.

Supplemental Material >

We explained the source of the power asymmetry in Section 2, and **Figure 4** illustrates it graphically, for the case of $F = 10$, $\beta = 0$, and $\rho = 0.6$. Panel *a* shows that $se(\hat{\beta}_{2SLS})$ is smaller in samples in which $\hat{cov}(z, u)$, the sample covariance between the instrument and the structural error, is greater. Panel *b* shows that $\hat{\beta}_{2SLS}$ is larger in such samples. The combination of these two forces causes $se(\hat{\beta}_{2SLS})$ to be smaller when $\hat{\beta}_{2SLS}$ is shifted toward OLS.

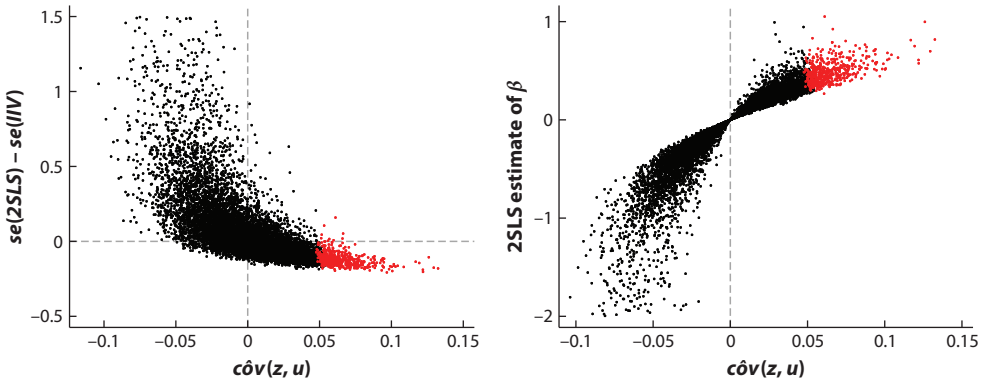


**Figure 4**

The figure shows $se(2SLS) - se(IIV)$ and $\hat{\beta}_{2SLS}$ plotted against $\hat{cov}(z, u)$, with $F = 10$ and $\rho = 0.6$. The red dots indicate $\hat{\beta}_{2SLS}$ significant at 5% level. Runs with $se(2SLS) - se(IIV) > 1.5$ or $\hat{\beta}_{2SLS} > 2$ or $< -2$ are not shown. Abbreviations: 2SLS, two-stage least squares; IIV, infeasible instrumental variables.

The DGP in **Figure 4** has $cov(z, x) = 0.10$. Hence, sample realizations of $\hat{cov}(z, u)$ that are large relative to 0.10 (i.e., roughly 0.05 or above) generate values of $\hat{\beta}_{2SLS}$ that are heavily shifted toward OLS, with spuriously small standard errors that are well below $se(IIV)$.

**Table 2** gives a broader view of the power asymmetry by showing Spearman's $r_s$ between $se(\hat{\beta}_{2SLS})$ and $\hat{\beta}_{2SLS}$ for different levels of $F$ and $\rho$. The table reveals how the negative association between 2SLS estimates and standard errors gets stronger as $\rho$ increases.

We can now understand why $t$-test rejection rates increase with $\rho$ in **Figure 1**. At low levels of endogeneity, the $t$-test has very low power unless instruments are quite strong, so it rejects $H_0{:}\beta = 0$ at rates well below 5%; but at higher levels of endogeneity, the $t$-test starts to get substantial spurious power from the finite sample correlation between $z$ and $u$, which causes estimates shifted in the direction of OLS to have low standard errors. As $\rho$ increases, the $t$-test derives more power from this source, causing it to reject $H_0{:}\beta = 0$ more frequently. Size inflation only appears when $\rho > 0.6$, as below that the $t$-test is underpowered.

We now examine the power of the 2SLS $t$-test in more detail by simulating the probabilities of rejecting $H_0{:}\beta = 0$ when it is false. We consider cases in which true $\beta$ is set to 0.30 or $-0.30$ in Equation 12. Importantly, these would be quantitatively large but plausible values in typical empirical applications, as they imply that a 1 standard deviation change in $x$ induces roughly a 0.30 standard deviation change in $y$. The results are reported in **Table 3**.

A key result is that the 2SLS $t$-test has almost no power to detect a sizeable true negative effect if the OLS bias is positive, unless the instrument is very strong. For example, as we see in **Table 3**, in the $F = 2.3$ case, widely viewed as an acceptable level of instrument strength, the probability of rejecting the false null $H_0{:}\beta = 0$ is only 0.2% when true $\beta$ is $-0.3$ and $\rho = 0.50$. Increasing instrument strength to $F = 10$ only increases power to 2.3%. The power asymmetry (the negative

**Table 2 Spearman rank correlations ($r_s$) between $se(\beta_{2SLS})$ and $\hat{\beta}_{2SLS}$**

| Population $F$ | $F_{5\%}$ | Spearman correlations ($r_s$) | | |
|---|---|---|---|---|
| | | $\rho = 0.2$ | $\rho = 0.5$ | $\rho = 0.8$ |
| 2.3 | 10 | −0.133 | −0.359 | −0.576 |
| 10 | 23.1 | −0.310 | −0.687 | −0.916 |
| 73.75 | 104.7 | −0.350 | −0.720 | −0.917 |

**Table 3** Power of 2SLS *t*-test: rejection of false null hypothesis $H_0$:$\beta = 0$ (%)

| Population $F$ | $F_{5\%}$ | $\beta = 0.3$ | | | $\beta = -0.3$ | | |
|---|---|---|---|---|---|---|---|
| | | $\rho = 0$ | $\rho = 0.5$ | $\rho = 1$ | $\rho = 0$ | $\rho = 0.5$ | $\rho = 1$ |
| 2.30 | 10.00 | 2.4 | 13.0 | 25.1 | 2.2 | 0.2 | 3.2 |
| 5.78 | 16.38 | 7.2 | 18.8 | 26.3 | 7.2 | 0.5 | 0.8 |
| 10.00 | 23.10 | 13.4 | 23.7 | 28.9 | 13.3 | 2.3 | 0.2 |
| 73.75 | 104.7 | 71.4 | 67.8 | 65.1 | 71.9 | 78.0 | 89.1 |

association between 2SLS estimates and standard errors) drives this result, as negative estimates have inflated standard errors if the OLS bias is positive.

We emphasize that if an instrument is strong enough to pass the Stock & Yogo's tests it is safe to assume that size inflation in two-tailed 2SLS *t*-tests is modest, and that 2SLS is approximately median unbiased. As we show in **Supplemental Appendix F**, for all values of $F$ considered in this section, median bias in 2SLS is negligible, or at least modest, regardless of the degree of endogeneity. Furthermore, size inflation was not a problem in any simulation in this section.

Supplemental Material >

The power asymmetry is a completely separate problem from bias and size inflation, and it remains serious at much higher levels of instrument strength: 2SLS estimates appear spuriously precise in samples where they are most shifted in the direction of the OLS bias. In the next section we describe a way to deal with this problem.

## 5. THE ANDERSON-RUBIN TEST

Fortunately, there is a simple solution to the problematic properties of the 2SLS *t*-test, which is to use the AR test instead (Anderson & Rubin 1949). In the single instrument case, the AR test of $H_0$:$\beta = 0$ is simply the *t*-test from the regression of $y$ on $z\hat{\pi}$, where $\hat{\pi}$ is the first-stage estimate of $\pi$. This is simply the second stage of 2SLS run by hand, so of course it generates the same estimate $\hat{\beta}_{2SLS}$, but the standard error of this regression is

$$se(AR) = \frac{\hat{\sigma}_{AR}}{\sqrt{TSS_{x,z}}} \text{ where } \hat{\sigma}_{AR} = \sqrt{(N-2)^{-1}\sum_i (y_i - (z_i\hat{\pi})\hat{\beta}_{2SLS})^2}. \quad 13.$$

The AR test is simply $t_{AR} = \hat{\beta}_{2SLS}/se(AR)$. Comparing Equations 13 and 9, we see the only difference is that $\hat{\sigma}_{AR}$ replaces $\hat{\sigma}_{2SLS}$. As we explained in Section 2, an important source of the 2SLS *t*-test power asymmetry is that $\hat{\sigma}_{2SLS}$ is a quadratic function of $\hat{\beta}_{2SLS}$ that is minimized at $\hat{\beta}_{OLS}$. In contrast, $\hat{\sigma}_{AR}$ is minimized at $\hat{\beta}_{2SLS}$, and its magnitude does not depend on the distance of $\hat{\beta}_{2SLS}$ from $\hat{\beta}_{OLS}$, as we show in **Supplemental Appendix C**. Thus, replacing $\hat{\sigma}_{2SLS}$ with $\hat{\sigma}_{AR}$ removes an important source of the power asymmetry.

We illustrate this in **Figure 5**, which compares 2SLS and AR standard errors in the case of $F = 10$, $\beta = 0$, and $\rho = 0.6$. Notice there is still a negative association between $se(AR)$ and $\hat{\beta}_{2SLS}$, which arises because $TSS_{x,z}$ tends to be larger when $\hat{\beta}_{2SLS}$ is close to OLS (see Section 2). However, the association is much weaker than for the 2SLS standard error. As a result, the AR test is far less subject to the power asymmetry problem: It generates rejections of the true null $\beta = 0$ roughly symmetrically distributed on the positive and negative side.

**Table 4** compares how frequently the *t*-test and AR test reject a true null hypothesis $H_0$:$\beta = 0$ at different levels of instrument strength. We again set $\rho = 0.60$, which is the best case for two-tailed *t*-test size distortion.[13] If $F = 10$, a 5% two-tailed *t*-test rejects $H_0$:$\beta = 0$ at a

[13] In **Figure 1** we see that *t*-test size is close to 5% when $\rho = 0.6$ regardless of instrument strength. At lower levels of $\rho$ size falls below 5%, and at higher levels of $\rho$ size is inflated. If the instrument is weak $\rho = 0.6$ is

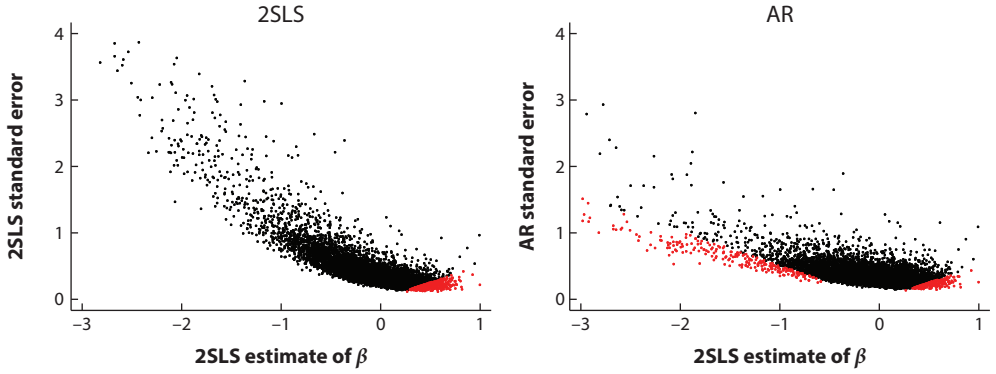


**Figure 5**

2SLS and AR standard errors plotted against $\hat{\beta}_{2SLS}$. Panel *a* shows 2SLS *t*-test rejections, and panel *b* shows AR test rejections. In each case, true $\beta = 0$, $F = 10$, and $\rho = 0.6$. The red dots indicate $H_0{:}\beta = 0$ rejected at 5% level. Runs with standard error >4 are not shown. Abbreviations: 2SLS, two-stage least squares; AR, Anderson-Rubin.

5.3% rate, so size distortion is trivial. However, all rejections occur when $\hat{\beta}_{2SLS} > 0$. Therefore, the size of a one-tailed 2.5% level *t*-test of $H_0{:}\beta \leq 0$ is 5.3%. Due to the power asymmetry, size distortions in one-tailed *t*-tests are large even with strong instruments. The mean significant $\hat{\beta}_{2SLS}$ is 0.483, as the significant estimates tend to be close to OLS. The *t*-test power asymmetry is severe even if we increase instrument strength to the high level of $F = 74$. Then size is 4.7%, but 83% of rejections occur when $\hat{\beta}_{2SLS} > 0$. The power asymmetry is not a weak instrument problem.

In contrast, the AR test rejects at exactly a 5% rate at all levels of instrument strength, and as long as the first stage *F* is at least 10, almost exactly half of those rejections occur when the estimate is positive. In the $F = 2.3$ ($F_{5\%} = 10$) case, the AR test also exhibits a power asymmetry, as 70% of rejections are positive. However, this is much milder than for the *t*-test, and the problem vanishes quickly as instrument strength improves.

Consider the implication of these patterns if there is publication bias, so that only estimates significant at the 5% level are published. If researchers rely on the *t* test, it will appear as if there is a consensus that $\beta$ is positive; but if researchers rely on the AR test, roughly half of studies will conclude $\beta$ is positive and half will conclude it is negative, and the literature as a whole will correctly conclude that there is no clear evidence that $\beta$ is nonzero.

**Figure 6** illustrates these ideas graphically. Panels *a* and *b* show the density of rejections of the true hypothesis $\beta = 0$ using the *t*-test (*black line*) versus the AR test (*red line*). We set $\rho = 0.60$ so that $E(\hat{\beta}_{OLS}) = 0.60$. As we see, essentially all *t*-test rejections occur when the 2SLS estimate

**Table 4 Anderson-Rubin (AR) versus *t*-test rejection rates, $H_0{:}\beta = 0$, case of $\rho = 0.6$**

| Population *F* | $F_{5\%}$ | *t* rejects, $\hat{\beta}_{2SLS} > 0$ | *t* rejects, $\hat{\beta}_{2SLS} < 0$ | $E(\hat{\beta}_{2SLS})$ if *t* rejects | AR rejects, $\hat{\beta}_{2SLS} > 0$ | AR rejects, $\hat{\beta}_{2SLS} < 0$ |
|---|---|---|---|---|---|---|
| 2.30 | 10.0 | 0.051 | 0.000 | 0.741 | 0.036 | 0.013 |
| 10.00 | 23.1 | 0.053 | 0.000 | 0.483 | 0.025 | 0.025 |
| 29.44 | 50.0 | 0.045 | 0.001 | 0.303 | 0.025 | 0.025 |
| 73.75 | 104.7 | 0.039 | 0.008 | 0.114 | 0.025 | 0.025 |

the Goldilocks level of endogeneity, so the *t*-test derives just enough spurious power from sample correlation between the *z* and *u* to reject $H_0{:}\beta = 0$ at about a 5% rate.

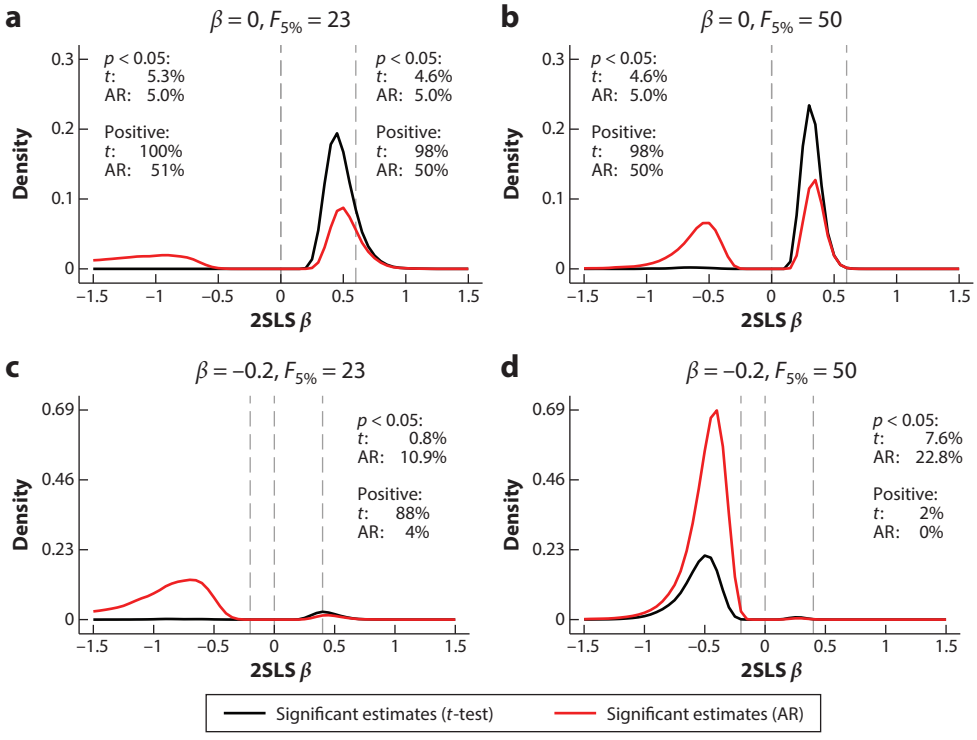


**Figure 6**

Frequency of rejecting $H_0$:$\beta = 0$ as a function of $\hat{\beta}_{2SLS}$. (*a*) $F = 10$ ($F_{5\%} = 23$), true $\beta = 0$, and the OLS bias is $\beta_{OLS} = 0.6$. (*b*) $F = 29.4$ ($F_{5\%} = 50$), true $\beta = 0$, and the OLS bias is $\beta_{OLS} = 0.6$. (*c*) $F = 10$ ($F_{5\%} = 23$), true $\beta = -0.20$, and $E(\hat{\beta}_{OLS}) = 0.40$. (*d*) $F = 29.4$ ($F_{5\%} = 50$), true $\beta = -0.20$, and $E(\hat{\beta}_{OLS}) = 0.40$. Abbreviations: 2SLS, two-stage least squares; AR, Anderson-Rubin; OLS, ordinary least squares.

is positive (in the direction of the OLS bias), even in the strong instrument case of $F = 29.4$ ($F_{5\%} = 50$). In contrast, the AR test generates balanced positive and negative rejections.

Panels *c* and *d* of **Figure 6** show what happens if we set true $\beta = -0.20$. This is a large effect in typical empirical applications, as it means that a 1 standard deviation change in $x$ induces roughly a $-0.20$ standard deviation change in $y$. The striking result is that the $t$-test has essentially no power to detect a true negative effect of this magnitude in the $F = 10$ ($F_{5\%} = 23$) case. Even in the quite strong instrument case of $F = 29.4$ ($F_{5\%} = 50$), it only detects a significant negative effect in 7.6% of cases, which is hardly informative as this scarcely exceeds the 5% size of the test. In contrast, AR detects a significant negative effect in 22.8% of cases. Therefore, the AR test has far better power to detect true negative effects than the $t$-test.

The practical implication of this result 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 strong by conventional standards, the 2SLS $t$-test will have little power to detect true negative effects. Thus, the use of the $t$-test in policy evaluation contravenes the principle of *primum non nocere* [first, do no harm].

The usual suggestion of the weak IV literature is to use AR in lieu of the $t$-test when the instrument is weak, because the AR test is robust—i.e., it has correct size regardless of instrument strength. However, we argue that the power asymmetry that plagues the $t$-test is a serious problem

even when instruments are strong. Hence we argue that the $t$-test should be abandoned in IV applications generally, and that AR should be used even with strong instruments.

Next, we give some general observations on the AR test. First, it is very easy to implement. In the one instrument case, the AR test of $H_0{:}\beta = 0$ is simply the $t$-test obtained by running 2SLS by hand and testing for significance of the fitted endogenous variable $\hat{x}$ in the second stage. Ironically, students have been taught to avoid doing this for years (e.g., Angrist & Pischke 2008, p. 140; Woolridge 2012, p. 529).[14]

Equivalently, the AR test is the $t$-test from the reduced-form regression of $y$ on the instrument $z$ and the exogenous controls. This is identical to the $t$-test from regressing $y$ on $z\hat{\pi}$ by hand, as multiplication by the scalar $\hat{\pi}$ does not alter the statistical significance of $z$.[15]

Second, in the one instrument case, the AR test is also equivalent to the $t$-test from the regression of $y$ on $\pi z$, the IIV estimator one could construct if one knew $\pi$. This may seem remarkable, but it stems from the fact that regressions of $y$ on $z$, $\pi z$, and $\hat{\pi} z$ yield identical $t$-stats. Therefore, AR is optimal and there is no good argument for using any other test. Moreira (2009) shows that AR is the uniformly most powerful unbiased test.

Third, with heteroskedastic or clustered data one should implement AR using a robust $t$-test to check for the significance of $z\hat{\pi}$ in the second stage of 2SLS (or, equivalently, of $z$ in the reduced-form regression of $y$ on $z$). Again, the AR test is not harder to implement than a conventional $t$-test, as one uses the exact same methods to obtain a robust test. In **Supplemental Appendix D** we show that our conclusions about the superior performance of the AR test carry over to DGPs with hetereoskedasticity/clustering. Moreira & Moreira (2019) show that the optimality of AR carries over to these settings.

Supplemental Material >

Fourth, one can use the AR test to form optimal confidence intervals. The standard error $se(AR)$ obtained by regressing $y$ on $z\hat{\pi}$ by hand does not give a valid confidence interval, and neither does the "correct" 2SLS standard error $se(\hat{\beta}_{2SLS})$ calculated by Stata. The distribution of $\hat{\beta}_{2SLS}$ is nonnormal in finite samples, meaning that valid confidence intervals cannot be symmetric. The correct procedure is called inverting the AR test, and it can be described as follows.

Assume that $\hat{\beta}_{2SLS} > 0$. Regress $y - b_L\hat{x}$ on $\hat{x}$ and find the smallest value of $b_L$ that renders $\hat{x}$ insignificant at the 5% level. This $b_L$ is the lower bound of the 95% AR confidence interval. If $b_L > 0$, the estimate is significantly different from zero at the 5% level. Next, regress $y - b_U\hat{x}$ on $\hat{x}$ and find the largest value of $b_U$ such that $\hat{x}$ is insignificant at the 5% level. This gives the upper bound of the 95% confidence interval. Section 8 shows how to do this in Stata.

If the instrument is very weak, so that first-stage $\hat{F} < 3.84$, then a 95% AR confidence interval for $\beta$ is unbounded. As Dufour (2004) notes, this is not a problem with AR but rather an accurate reflection of uncertainty. If $\hat{F} < 3.84$, we do not have 95% confidence that the instrument is significant in the first stage, so we lack 95% confidence that the model is identified. It is an odd property of the $t$-test that it gives a bounded confidence interval in this case. In fact, if your first stage $\hat{F}$ is that small you should not be running 2SLS anyway!

[14]For example, Woolridge (2012, p. 529) states: "You should avoid doing the second stage manually, as the standard errors and test statistics obtained in this way are not valid." Instead, students are advised to let Stata construct the 2SLS residuals $\hat{u} = y - x\hat{\beta}_{2SLS}$ and use these to estimate the asymptotic variance of $\hat{\beta}_{2SLS}$, as in Equation 9. This gives asymptotically correct standard errors, which have strange properties in finite samples (as we have seen). It is worth noting that $\hat{u}$ is not a prediction error like an OLS residual, as we would never predict $y$ by multiplying endogenous $x$ by $\hat{\beta}_{2SLS}$, which is designed to capture the impact of exogenous changes in $x$. That is why an $R^2$ based on these residuals is meaningless.

[15]Recall that the reduced from is $y = z(\beta\pi) + (\beta e + u) = z\xi + v$, where $\xi = \beta\pi$. The AR test judges $\hat{\beta}_{2SLS}$ to be significant if $z$ is a significant predictor of $y$ in this regression. A valid instrument $z$ must satisfy $\pi \neq 0$ and $cov(z, v) = 0$, so a test of the hypothesis $H_0{:}\xi = 0$ is also a test of $H_0{:}\beta = 0$.

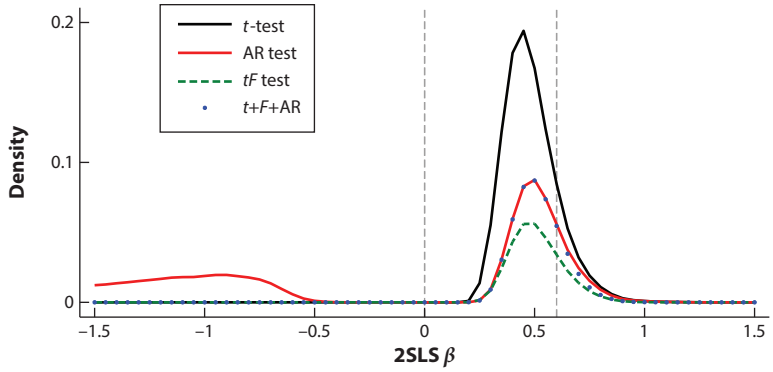


**Figure 7**

Frequency of rejecting $H_0$:$\beta = 0$ as a function of $\hat{\beta}_{2SLS}$, with $F = 10$, $\beta = 0$, and $E(\beta_{OLS}) = 0.6$.
Abbreviations: 2SLS, two-stage least squares; AR, Anderson-Rubin; OLS, ordinary least squares.

Fifth, with multiple instruments the AR test of $H_0$:$\beta = 0$ is simply the joint $F$ test for significance of the vector of instruments in the reduced form. However, it is not optimal. Instead, Moreira (2003) shows that the CLR test is the uniformly most powerful unbiased test (given homoskedastic data). The AR and CLR tests are equivalent in the single instrument case. Finlay & Magnusson (2009) provide a Stata command for a heteroskedasticity-robust version of CLR developed by Kleibergen (2005) and to invert it to form confidence intervals. We explain CLR in detail in other work (Keane & Neal 2023), but here we continue to focus on the single instrument case in the interest of expositional simplicity.

Finally, we consider some proposed alternatives to the AR test. A fair summary of the advice commonly given to applied researchers is to view 2SLS results as reliable only if (*a*) $\hat{F} > 10$, (*b*) the *t*-test is significant, and (*c*) the AR test is also significant, providing a robustness check (see Angrist & Pischke 2008, p. 212). In **Figure 7** we compare this procedure, which we denote *t*+F+AR, to simply relying on the AR or *t*-test. These results are for the case of $F = 10$, $\beta = 0$, and $\rho = 0.60$. As before, the red line plots the density of estimates that are significant according to AR, and the black line plots the *t*-test results. The blue dotted line shows the $t + F$+AR test results. This procedure suffers from a severe power asymmetry, similar to the *t*-test. It succeeds in reducing the frequency of positive rejections, as unlike the *t*-test it does not suffer from inflated power to detect false positives, but it has no better power on the negative side than the *t*-test.

Recently, Lee et al. (2022) have advocated the *tF* test, which increases the *t*-test critical value to eliminate worst-case size distortion. This is shown by the green dotted line. As we see, the *tF* test has very low power and does not solve the power asymmetry problem. **Figure 1** shows how *t*-test power depends on both $F$ and $\rho$. Lee et al. (2023) advocate a *VtF* test, which increases or reduces *t*-test critical values based on $\hat{F}$ and $\hat{\rho}$, so size is always 5%. However, this adjustment also fails to address the power asymmetry problem.

Finally, as we see in **Figure 1**, *t*-test size inflation does not arise unless endogeneity is strong ($\rho > 0.60$). Angrist & Kolesár (2024) have defended the *t*-test on this basis. We argue that the results in **Figure 1** are not comforting, as they reflect the very poor power properties of the *t*-test, which the AR test avoids.

## 6. PERFORMANCE OF 2SLS RELATIVE TO OLS

Weak instrument tests focus on bias and size distortions in two-tailed 2SLS *t*-tests. However, applied researchers use 2SLS because they expect it to deliver more reliable estimates than OLS.

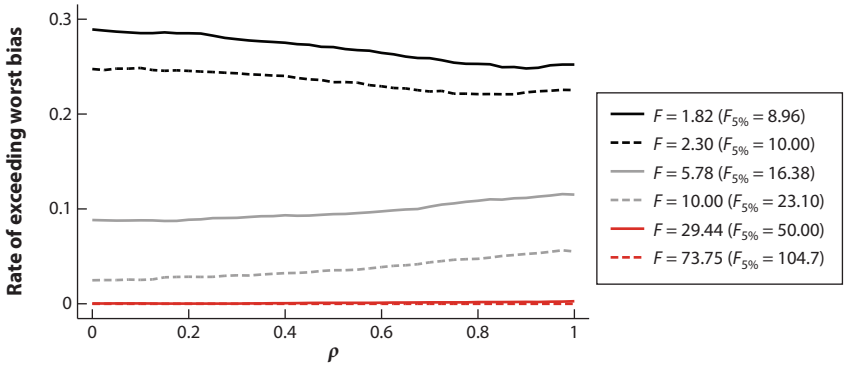


**Figure 8**

Probability that the 2SLS estimation error exceeds the worst-case OLS bias. The figure shows the % of runs with $|\hat{\beta}_{2SLS}| > 1$. Abbreviations: 2SLS, two-stage least squares; OLS, ordinary least squares.

So we ask, How strong must instruments be to give high confidence that 2SLS will give more reliable estimates than OLS? If instruments fail to meet this minimal standard, one would be well advised to find better instruments or consider alternative approaches.

We start to explore this question by asking how often 2SLS estimation errors exceed the worst-case OLS bias that arises if $x$ is perfectly correlated with the error in the outcome equation. To do this, we simulate data sets from the model in Equation 12, setting true $\beta = 0$, and count the frequency of estimation runs that generate extreme outliers where $|\hat{\beta}_{2SLS}| > 1$.[16] As we see in **Figure 8**, the risk of such extraordinarily large outliers is a remarkable 22–25% in the case of $F = 2.30$ ($F_{5\%} = 10$), which corresponds to Staiger & Stock's rule of thumb for acceptable instrument strength. A much stronger instrument level of $F = 29.44$ ($F_{5\%} = 50$) is needed to render such large outliers virtually impossible.

**Figure 9** plots the density of 2SLS estimates in the case of $\rho = 0.6$ for three levels of instrument strength. In the $F = 2.3$ ($F_{5\%} = 10$) case, which corresponds to Staiger & Stock's rule of thumb, the distribution is highly nonnormal, with fat tails, high frequency of extreme outliers, and left skewness all very apparent. Only by increasing instrument strength to the much higher level of $F = 29.44$ ($F_{5\%} = 50$) does normality appear to be a decent approximation to the sampling distribution of the 2SLS estimator.

**Figure 9** also plots the mean OLS estimate, 0.60, and the 95% OLS confidence interval. Careful inspection of the figure reveals that, due to their high dispersion, the 2SLS estimates are frequently farther from the true value ($\beta = 0$) than the OLS estimates. This is especially true in the $F_{5\%} = 10$ and $F_{5\%} = 23.1$ cases, but it is much less common in the $F_{5\%} = 50$ case.

Next, in **Figure 10**, we report the fraction of simulated data sets where 2SLS performs worse than OLS, meaning that the 2SLS estimate is farther from the true $\beta$ than OLS. Of course, at low levels of endogeneity ($\rho \approx 0$) OLS almost always wins, 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 $\rho$. Take the case of $C = 2.30$ ($F_{5\%} = 10$), which corresponds to Staiger & Stock's rule of thumb. The value of $\rho$ has to approach 0.50 before the probability that 2SLS outperforms OLS passes 50%. In this context, it is worth recalling that $\rho = 1$ is the highest possible level of endogeneity, so $\rho = 0.5$ corresponds to a fairly high level of endogeneity.

[16] We set $N = 1{,}000$, but recall that $N$ is not very important, as finite sample properties of 2SLS do not depend on $N$ or first-stage $R^2$ per se but only on how they combine to form $F = NR^2/(1 - R^2)$.

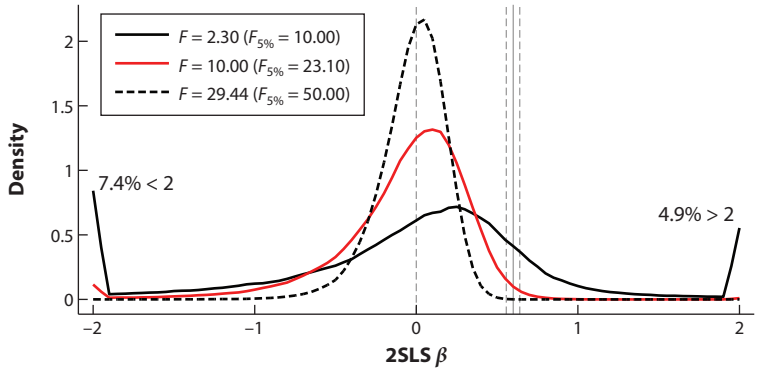


**Figure 9**

Kernel density of 2SLS estimates when $\rho = 0.6$ and estimates are censored to $\pm 2$. The figure is based on 100,000 simulated data sets of size $N = 1{,}000$. Densities are estimated using a bandwidth of 0.03. The gray line shows the mean OLS estimate, while the nearby dotted lines are the 95% CI. Abbreviations: 2SLS, two-stage least squares; CI, confidence interval; OLS, ordinary least squares.

In many applications, one can put a reasonable prior on $\rho$ (i.e., the OLS bias) and assess the performance of 2SLS relative to OLS for different levels of instrument strength in that context. For example, consider the archetypal application of IV to estimating a regression of log wages on education. Using Panel Study of Income Dynamics (PSID) data from 2015, we calculate a correlation between education and log earnings of 0.45.[17] 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 $\rho$ is 0.45. Therefore, in such an application a uniform prior on $\rho \in [0, 0.45]$ may be reasonable.[18]

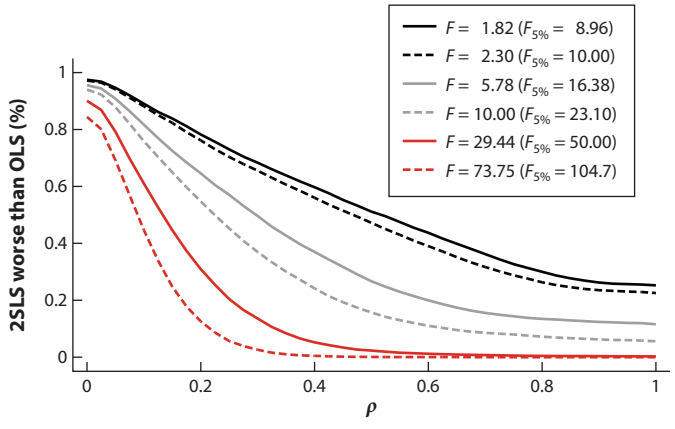


**Figure 10**

Probability of 2SLS performing worse than OLS. The figure plots the proportion of Monte Carlo replications where $|\hat{\beta}_{2SLS} - \beta| > |\hat{\beta}_{OLS} - \beta|$. Abbreviations: 2SLS, two-stage least squares; OLS, ordinary least squares.

[17]We use data on household heads aged 30–54 and partial out the effects of age and age$^2$. The wage is constructed as labor income/hours. We screen on hours $\in$ [400, 4,160], income $\in$ [\$3,000, \$235,884], wage >2.70 per hour, and valid data on education and labor income. This gives $N = 3{,}634$.

[18]If education is measured with error, $\rho$ may go negative. We discuss a measurement error DGP in **Supplemental Appendix D**. **Figure 10** is symmetric about $\rho = 0$, so we do not show the negative side. Allowing $\rho < 0$ would make 2SLS look worse than in our calculations, as it puts more mass near $\rho = 0$.

Supplemental Material >

**Table 5 Probability of 2SLS outperforming OLS**

| Population $F$ | $F_{5\%}$ | Uniform prior for $\rho$: | | | |
|---|---|---|---|---|---|
| | | 0 to 1 | 0 to 0.45 | 0.35 to 0.45 | 0.5 to 1 |
| 1.82 | 8.96 | 45 | 24 | 41 | 65 |
| 2.30 | 10 | 48 | 26 | 45 | 69 |
| 3.84 | 13 | 56 | 32 | 55 | 77 |
| 5.78 | 16.38 | 62 | 38 | 64 | 84 |
| 10.00 | 23.10 | 70 | 47 | 77 | 91 |
| 29.44 | 50 | 83 | 65 | 95 | 99 |
| 73.75 | 104.70 | 89 | 76 | 100 | 100 |

We report the frequency of $|\hat{\beta}_{2SLS} - \beta| < |\hat{\beta}_{OLS} - \beta|$ across Monte Carlo replications, averaged across all values of $\rho$ under a uniform prior that $\rho$ falls in the indicated range. Abbreviations: 2SLS, two-stage least squares; OLS, ordinary least squares.

Applied researchers may find a prior on $\rho$ unfamiliar, so it is worth noting that plausible values of $\rho$ can be backed out empirically given any hypothesized value of $\beta$. For $\beta = \beta^p$, the implied value of $\rho$ is the correlation of the residuals from (*a*) the regression of $y - x\beta^p$ on $z$ and (*b*) the first-stage regression of $x$ on $z$.[19] For example, if $\beta^p = 0$ this is simply the correlation of the reduced-form errors, and if $\beta^p = \beta_{OLS}$ this is zero. Thus, a prior that $\beta \in (0, \beta_{OLS})$, which may be natural in many applications where one suspects positive selection into treatment, corresponds to a prior that $\rho$ lies between zero and the correlation of the reduced-form residuals. This is how we motivate the uniform prior on $\rho \in [0, 0.45]$ in the example of wages and education.

**Table 5** shows the probability that 2SLS will outperform OLS given different levels of instrument strength and different priors on $\rho$. For example, if $F = 2.30$ ($F_{5\%} = 10$), which corresponds to Staiger & Stock's rule of thumb for strong instruments, and given a uniform prior $\rho \in [0, 0.45]$, the probability that 2SLS outperforms OLS is only 26%.

Alternatively, a researcher who thinks education is very highly correlated with ability might have a uniform prior of $\rho \in [0.35, 0.45]$. Even then, the probability that 2SLS beats OLS is only 45%. Clearly, in an application to estimating the effect of education on earnings, one should require a substantially higher level of instrument strength than the $\hat{F} \geq 10$ rule of thumb, even if one believes endogeneity is very severe.

If we increase instrument strength to the $F = 29.44$ ($F_{5\%} = 50$) level, a uniform prior on $\rho \in [0, 0.45]$ implies a 65% chance that 2SLS will outperform OLS. Even that level of performance is not too inspiring. If we have a uniform prior on $\rho \in [0.35, 0.45]$, meaning we think endogeneity is very severe, the probability of 2SLS beating OLS increases to 95%.

Thus, in the archetypal application of IV to estimating the return to education, one clearly needs an $\hat{F}$ of at least 50 to have high confidence that 2SLS will outperform OLS. Even then, one's confidence does not reach 95% unless one believes ability bias is severe.

## 6.1. Practical Advice on Acceptable First-Stage *F* Levels

In general, the results of this section suggest that instruments should be much stronger than standard thresholds, like the popular $\hat{F} > 10$, to give confidence that 2SLS results are likely to be superior to OLS, in the sense that $|\hat{\beta}_{2SLS} - \beta| < |\hat{\beta}_{OLS} - \beta|$. The level of instrument strength required to have confidence that 2SLS will outperform OLS depends heavily on one's prior about $\rho$. We suggest that researchers should assess the level of instrument strength required to

[19] Of course one should also include any exogenous control variables present in the application. For example, in our wage example we control for age and age$^2$.

have reasonable confidence that 2SLS will outperform OLS in any particular application, based on reasonable priors on the severity of the endogeneity problem ($\rho$).

Despite the difficulty of devising a general rule of thumb, a strong case can be made that applied researchers should adopt a threshold of instrument strength of at least $\hat{F} > 50$ in the single instrument case. This makes 2SLS likely to outperform OLS at moderate levels of endogeneity, although at high levels of $\rho$ a lower $\hat{F}$ would suffice.[20] As we have seen, this threshold renders extreme 2SLS outliers very unlikely. If such a threshold cannot be met, it is advisable to seek stronger instruments or pursue alternative strategies, such as OLS combined with a serious attempt to control for omitted variables.[21] We reiterate that robust tests (AR or CLR) should be used in lieu of 2SLS $t$-tests regardless of $\hat{F}$.

Our discussion of properties of the 2SLS estimator has focused on the independent and identically distributed (iid) normal model of Equation 12 to emphasize key issues.[22] However, in assessing acceptable first-stage $F$ statistics in practice, it is important to consider the impact of heteroskedasticity and clustering and to use a heteroskedasticity- or cluster-robust $F$ statistic, as we discuss in Section 8.

### 6.2. Why Does 2SLS Perform So Poorly Relative to OLS?

The focus of Section 6 on the quality of estimates may seem unrelated to the focus of Sections 2 through 5 on test statistics. Conceptually, however, the reason 2SLS can perform so poorly relative to OLS even when an instrument is quite strong by conventional standards is closely related to the issue of the magnitude of $\hat{cov}(z, u)$ relative to $cov(z, x)$ that we discussed in Section 2.

Consider the strong instrument case of $F = 10$ ($F_{5\%} = 23.1$). Given a sample size of $N = 1{,}000$, this corresponds to a population correlation between the instrument and endogenous variable of $corr(z, x) = 0.10$. At this level, even small sample realizations of $\hat{corr}(z, u)$ can drive the 2SLS estimate far from the true value. This is clear from an inspection of the right panel of **Figure 4**. While the use of IV solves the endogeneity problem that arises because $cov(x, u) \neq 0$, it worsens by an order of magnitude the sampling problem created by the fact that $\hat{cov}(z, u) \neq 0$ in finite samples.

Increasing sample size does not solve this problem for the following reason. Assume that $N = 100{,}000$. Then, $F = 10$ corresponds to a population correlation between the instrument and endogenous variable of only $corr(z, x) = 0.01$. We can expect the larger sample size to reduce $\hat{corr}(z, u)$ by an order of magnitude, but it is no smaller compared to $corr(z, x)$!

## 7. PRACTICAL RELEVANCE IN SELECTED INSTRUMENTAL VARIABLES PAPERS IN THE *AMERICAN ECONOMIC REVIEW*, 2011–2013

We explore the practical importance of the issues we have discussed by examining results from IV papers published in the *American Economic Review* (AER) from January 2011 to August 2023. We searched for papers with a first-stage $F$ either below 50 or unknown, which relied on the $t$-test for inference. We identified 102 papers, of which 53 could not be replicated, usually due to confidential data. In the remaining 49, we examined whether $t$-test results are reversed by the use

[20] With $K > 1$ instruments, $\hat{F}$ contains a factor of $1/K$, so it falls mechanically with $K$. So an $\hat{F}$ smaller than 50 is needed to maintain the same probability of 2SLS outperforming OLS. We find that the required $\hat{F}$ is slightly larger than $50/K$; very roughly, it is about $50/K^{3/4}$.

[21] There are also alternatives to 2SLS that we evaluate in **Supplemental Appendix G**: the Fuller and JIVE estimators and the unbiased estimator of Andrews & Armstrong (2017). JIVE performs worse than 2SLS, but the Fuller and unbiased estimators do somewhat better if endogeneity is severe and instruments are weak. However, these improvements are not great enough to change our basic advice about acceptable first-stage $F$.

[22] This is less restrictive than it appears: Andrews et al. (2019) show that for any heteroskedastic DGP, there exists a homoskedastic DGP giving equivalent behavior of 2SLS estimates and test statistics.

Supplemental Material >

**Table 6 Problems with *t*-test inference in AER papers, 2011–2023**

| Article | Relevant table | First-stage *F*-statistic | IV versus OLS estimate | AR versus *t*-test *p*-value |
|---|---|---|---|---|
| **Just-identified IV models** | | | | |
| Alesina & Zhuravskaya (2011) | Table 6, column 4 | 8.4 | OLS = −1.2<br>IV = −1.8 | *t* = 0.023<br>AR = 0.092 |
| Autor et al. (2020) | Table 3, column 5 | 29.0 | OLS = 14.0<br>IV = 46.0 | *t* = 0.09<br>AR = 0.04 |
| Hornung (2014) | Table 5, column 3 | 5.7 | OLS = 1.59<br>IV = 1.67 | *t* = 0.05<br>AR = 0.093 |
| Juhász (2018) | Table 4, column 6 | 10.3 | OLS = 2.47<br>IV = 2.68 | *t* = 0.004<br>AR = 0.13 |
| Markevich & Zhuravskaya (2018) | Table 5, column 3 | 17.3 | OLS = 0.92<br>IV = 0.78 | *t* = 0.013<br>AR = 0.095 |
| Pascali (2017) | Table 6, column 3 | 25.0 | OLS = −0.05<br>IV = −0.19 | *t* = 0.045<br>AR = 0.083 |
| Rao (2019) | Table 3, column 4 | 72.0 | OLS = 4.59<br>IV = 8.40 | *t* = 0.036<br>AR = 0.093 |
| Shapiro & Walker (2018) | Table 1, column 1 | 14.0 | OLS = ?*<br>IV = 130.03 | *t* = 0.043<br>AR = 0.16 |
| **Overidentified IV models** | | | | |
| Autor & Dorn (2013) | Table 5, column 7 | 49.0<br>[*k* = 3] | OLS = 0.11<br>IV = 0.15 | *t* = 0.007<br>CLR = 0.119 |
| Fang & Gavazza (2011) | Table 2, column 2 | 19.2<br>[*k* = 6] | OLS = 0.03<br>IV = 0.51 | *t* = 0.013<br>CLR = 0.056 |
| Nakamura & Steinsson (2014) | Table 2 | **<br>[*k* = 50] | OLS ∼0.2–0.8<br>IV ∼1.3–1.9 | *t* ∼ 0.00–0.05<br>CLR ∼0.36–0.57 |
| Voors et al. (2012) | Table 4, column 7 | 6.0<br>[*k* = 2] | OLS = 0.06<br>IV = 0.07 | *t* = 0.052<br>CLR = 0.302 |

Heteroskedasticity- or cluster-robust versions of *t*, AR, and CLR tests are implemented consistent with each paper's methodology. Olea & Pflueger's (2013) first-stage *F* is reported for overidentified IV models. One asterisk indicates the paper cannot be replicated and OLS results are not reported; its inclusion is possible as the authors report both first-stage and reduced-form results. Two asterisks indicate the first-stage *F* cannot be calculated due to issues with the covariance matrix after clustering. Abbreviations: AER, *American Economic Review*; AR, Anderson-Rubin; CLR, conditional likelihood ratio; IV, instrumental variables; OLS, ordinary least squares.

of AR or of CLR in overidentified cases. We implemented heteroskedasticity- or cluster-robust versions of these tests, as appropriate for the data in each paper.[23] We found that at least one key result is overturned in 12 of the 49 papers (24%).

Supplemental Material >

These 12 articles are described in **Table 6** and **Supplemental Appendix H**. Eight use just-identified models, while 4 use overidentified models. In 11 of 12 cases, a result that is significant according to the *t*-test is rendered insignificant by using AR or CLR.[24] A clear pattern emerges where, in most of these cases, the 2SLS estimate is close to OLS. It is precisely in such cases that we expect the 2SLS standard error to be too small due to the power asymmetry problem, making the *t*-test overly likely to reject the null hypothesis.

[23] For the CLR test, we rely on Kleibergen's (2005) extension of the CLR to GMM.

[24] Many papers report more than one key result, so we are not necessarily saying that all results are reversed. We only considered main results and not robustness checks and so on.

The paper by Autor et al. (2020) is the sole instance where use of AR gives a stronger significance result than the *t*-test. In this case the 2SLS estimate is far from OLS, which is precisely when we expect the 2SLS standard error to be too large.

Given publication bias, we expect to find many more published papers where use of AR or CLR renders *t*-test results insignificant than vice versa. IV regressions that give estimates far from OLS tend to have spuriously large standard errors, making them less likely to be published in the first place. In other words, there is selection bias, as we cannot see papers that would have been published in AER if the AR or CLR test had been trusted for inference instead of the *t*-test and if these tests had given more significant results. Therefore, this pattern does not indicate that the AR or CLR tests are generally more conservative than the *t*-test.

## 8. A SIMPLE GUIDE FOR APPLIED RESEARCHERS

Here we present a simple guide for applied researchers seeking to implement IV inference while avoiding the problems with the 2SLS *t*-test. In the single instrument case we show how to use the AR test, and in the multiple instrument case we show how to use CLR. Given its widespread use among applied economists, we present Stata code to implement our suggestions. When we examined AER papers from 2011 to 2023, we found only 3 out of 52 that met our search criteria, used public data, and used software other than Stata.[25]

First, consider the case in which outcome $y$ is regressed on the single endogenous variable $x$ and an exogenous control variable $c$, and where $z$ is the excluded instrument. We suggest reporting results from the following procedures.

1. Run and report OLS. It is important to compare 2SLS and OLS results.
   In Stata: `reg y x c, vce(`*type*`)`
2. Run the first-stage regression of $x$ on $z$ and $c$.
   In Stata: `reg x z c, vce(`*type*`)`
3. Obtain a heteroskedasticity/cluster-robust $F$ statistic for significance of the instrument.
   In Stata: `test z = 0`
4. Compute $\hat{x}$.
   In Stata: `predict xhat, xb`
5. Run the second-stage regression of $y$ on $\hat{x}$ and use the $t$-test from this regression to test $H_0{:}\beta = 0$. This $t$-stat is the AR test of $H_0{:}\beta = 0$ in the one instrument case.
   In Stata: `reg y xhat c, vce(`*type*`)`
6. Construct a valid confidence interval by inverting the AR test.
   In Stata: `weakiv ivregress 2sls y (x = z) c, vce(`*type*`)`

In all these commands the "vce" option determines how the variance matrix of the parameter estimates is calculated. The results are rendered robust to heteroskedasticity or clustering via the option one specifies for "*type*," which refers to the data type—for example, vce(cluster personid) for panel data. Enter `help reg` in Stata for a list of options.

In step 3 it is important to use a robust $F$ statistic, as recent papers by Andrews et al. (2019) and Young (2022) emphasize 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. As we showed in Section 6, if first-stage sample $F$ is below roughly 50, one may be concerned that the 2SLS estimate may be no more reliable than OLS.

[25] Those three papers used Matlab and R. We did not replicate them as we lack sufficient familiarity with those software packages.

In step 5, the $t$-statistic and $p$-value from the second stage regression give the AR test of $H_0$:$\beta = 0$. The $F$ version of the AR test is the square of the $t$-statistic from the second-stage regression.[26] Always remember that the standard error from the-second stage regression is not valid for forming confidence intervals. Instead, the AR test should be inverted as in step 6, using the Stata command `weakiv` by Finlay et al. (2013).[27]

In step 6, if the confidence interval reported by `weakiv` is unbounded—i.e., it covers the entire grid—it means you do not have at least 95% confidence that the model is identified. This will happen if first-stage $F$ is less than 3.84. You should not be doing IV in this case!

Importantly, our results suggest that the above methodology is preferred for all first-stage $F$ statistics, not simply ones that are currently considered in the literature to be weak. As the strength of the instrument grows, inferences using the AR test and the $t$-test converge.

Finally, we note that the popular user command `ivreg2` in Stata also contains the AR test statistic, but we do not recommend it for two reasons. First, it does not (at the time of writing) report AR confidence intervals. Second, it reports 2SLS $t$-test results that we argue are better left unknown, as this may lead to researchers trying to screen on specifications that contain a significant $t$-test. This is undesirable for reasons shown in **Figure 7**.

Next we consider the case of multiple instruments. In this case, we strongly advise against using 2SLS or the two-step GMM estimator (GMM-2S), which extends 2SLS to heteroskedastic data, as both suffer from severe bias toward OLS. Furthermore, the associated $t$-tests suffer from severe size inflation and power asymmetry.

As we show elsewhere (Keane & Neal 2023), given homoskedasticity, the limited information maximum likelihood (LIML) estimator and associated CLR test largely avoid these problems. Continuously updated GMM, often called CUE, generalizes LIML to heteroskedastic data, while Kleibergen (2005) provides a generalization of CLR. We recommend using these procedures in the overidentified case.[28] For concreteness, consider a case with two instruments `z1` and `z2`. Then follow the procedure below.

1. Run and report OLS. It is important to compare 2SLS and OLS results.
   In Stata: `reg y x c, vce(`*type*`)`
2. Run the CUE estimator.[29] The $\hat{\beta}_{\text{CUE}}$ is only used to obtain the estimate of $\beta$. The CUE standard error and $t$-stat should not be used for inference.
   In Stata: `ivreg2 y (x = z1 z2) c, cue` *type*
3. For inference obtain Kleibergen's (2005) extension of CLR for GMM.
   In Stata: Run `weakiv` after step 2. The Kleibergen test of $H_0$:$\beta = 0$ will simply be called CLR.[30] A 95% confidence interval is also provided by default.[31]

[26] You may notice the AR test statistics reported in `weakiv` and `ivreg2` are different from the $t$-statistic in the manual second-stage regression. `ivreg2` reports both the $F$ and $\chi^2$ versions of the AR test, while `weakiv` reports the $\chi^2$ version only. These two versions can diverge when clustered standard errors are used with too few clusters.

[27] If an AR test of $H_0$:$\beta = \beta_0$ is desired, use the $t$-test from regressing $y - x\beta_0$ on $\hat{x}$ and $c$.

[28] As IV strength increases, LIML and CUE converge to 2SLS and two-step GMM, respectively.

[29] The CUE is obtained via an iterative procedure. If it fails to converge, or is too computationally burdensome, LIML with a robust variance matrix is a good fallback option. However, it is less efficient than CUE under heteroskedasticity.

[30] If a test of $H_0$:$\beta = \beta_0$ is desired, one may regress $y - x\beta_0$ on the instrument vector. The $F$-test from this regression is the AR test in the multiple instrument case, but it is less efficient than the CLR test obtained by replacing $y$ with $y - x\beta_0$ in step 2 and repeating step 3.

[31] Inverting the CLR test requires a grid search over $b_L$ and $b_U$. The command `weakiv` uses a default grid size, but the user can specify a finer grid if desired. This increases computation time.

4. Obtain Hansen's (1982) J-test. If it rejects the overidentifying restrictions, there is evidence the instruments are invalid so the results should not be trusted.
   In Stata: The J-test is provided in the `ivreg2` results (see step 2).
5. Obtain and report Olea & Pflueger's (2013) first-stage $F$ statistic.
   In Stata: Run `weakivtest` after step 2.[32]

There are some published papers that use 2SLS and the $t$-test but then report CLR as a robustness check. However, we advise against this procedure. First, it leads to the same type of problems we explained in **Figure 7**. Second, CLR is designed for use with LIML or CUE, and, as we explain elsewhere (Keane & Neal 2023, figure 10), mixing and matching estimators and test statistics based on different estimators can generate odd results.

We use Olea & Pflueger's (2013) first-stage $F$ statistic (O-F) because, as Andrews et al. (2019) note, it is inappropriate to use either a conventional or an heteroskedasticity-robust $F$ to gauge instrument strength in nonhomoskedastic overidentified settings.[33]

Lastly, it is worth noting that the CLR test is optimal in the overidentified case with homoskedastic errors. However, the theory literature has not settled on an optimal test for the heteroskedastic case. Therefore, there are some proposed alternatives to the Kleibergen test that we discuss in **Supplemental Appendix I**. In our experience these tests give similar results unless the data are poorly behaved (e.g., clustered data with few clusters).

Supplemental Material >

## 9. CONCLUSION

Since the work of Stock & Yogo (2005), the weak IV literature has focused heavily on the issue of size inflation in two-tailed 2SLS $t$-tests.[34] The Stock–Yogo tests indicate that a first-stage $\hat{F}$ in the 10–20 range is sufficient to guarantee that size inflation is modest in exactly identified models. However, we argue that the emphasis on size inflation of two-tailed $t$-tests has caused the literature to gloss over other important problems that persist even when instruments are much stronger.

In particular, 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 close to OLS or are strongly shifted in the direction of the OLS bias. This association between 2SLS estimates and standard errors, which we call the power asymmetry, persists even if instruments are very strong. This has two important consequences: 2SLS estimates that are close to OLS, or shifted strongly toward OLS, will appear spuriously precise, so the $t$-test has inflated power to judge such estimates significant; and conversely, 2SLS $t$-tests have little power to detect a true $\beta$ that is far from the OLS estimate.

Fortunately, one can largely avoid the power asymmetry problem by using the AR test of Anderson & Rubin (1949). Furthermore, the AR test has correct size even when instruments are weak. Hence, we argue that applied researchers should abandon the 2SLS $t$-test altogether and adopt AR instead, regardless of the level of instrument strength.

We illustrate the practical importance of the power asymmetry problem by examining IV papers published in AER from 2011–2023. We consider 49 replicable papers where the first stage $F$ is below 50 or unknown. In one quarter of these papers a key result obtained using the $t$-test is

[32] The critical values the command reports are for maximal bias, which is not the only concern.
[33] In the single instrument case, both O-F and Kleibergen & Paap's (2006) statistic reduce to the conventional robust $F$.
[34] Stock & Yogo (2005) also test bias relative to OLS. However, this criterion can only be assessed given overidentification of degree 2, which is less common in practice than exact identification.

overturned by using the AR test.[35] In particular, we see many cases where an IV estimate close to OLS is judged significant by the $t$-test but insignificant by AR. This is the pattern we expect to see given the power asymmetry problem.

Another important consequence of the power asymmetry is that size distortions in one-tailed $t$-tests are far greater than in two-tailed tests. We find that first-stage $F$ levels in the thousands are required to reduce size distortions in one-tailed $t$-tests to modest levels. The reason applied researchers rarely use one-tailed tests is that they think two-tailed tests are symmetric (so a two-tailed 5% test is equivalent to a one-tailed 2.5% test). It is important to understand this is not even close to being true for 2SLS, even with very strong instruments.

The weak IV literature's heavy focus on $t$-test size inflation has also deflected attention from the quality of 2SLS estimates. However, what applied researchers really want is for 2SLS to give estimates closer to the truth than OLS. Given typical weak IV test thresholds, we find substantial probabilities of 2SLS performing worse than OLS. For example, a first-stage $\hat{F} \geq 10$ is a test of whether population $F$ is at least 2.3. At this level of instrument strength, and given a uniform prior on the degree of the endogeneity, we calculate a 52% probability that 2SLS will generate an estimate of $\beta$ farther from the truth than OLS.

Thus, we advise applied researchers to adopt a higher standard of instrument strength in IV applications. A strong case can be made for a first-stage acceptable $\hat{F}$ threshold of at least 50, although lower values are acceptable if endogeneity is severe. At lower levels of instrument strength, 2SLS estimates are likely to be even farther from the truth than OLS.

We have focused primarily on the single instrument case to make our presentation as simple as possible for applied researchers, but elsewhere (Keane & Neal 2023, 2024) we discuss the multiple instrument case in detail. As we show, the use of multiple instruments makes the power asymmetry that plagues the $t$-test worse, and it makes the 2SLS bias and $t$-test size inflation problems much more severe. Fortunately, however, we find that the LIML estimator of Anderson & Rubin (1949), used in conjunction with the CLR test of Moreira (2003), avoids these problems. The continuously updated GMM (CUE) of Hansen et al. (1996) and the CLR test of Kleibergen (2005) extend these procedures to heteroskedastic data. In our view these are the best choices in the overidentified case.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

We thank an anonymous referee, along with Isaiah Andrews, Josh Angrist, Michel Kolesar, Leandro Magnusson, Robert Moffitt, and Peter Phillips for valuable comments. This research was supported by Australian Research Council grants DP210103319 and CE170100005.

## LITERATURE CITED

Alesina A, Zhuravskaya E. 2011. Segregation and the quality of government in a cross section of countries. *Am. Econ. Rev.* 101(5):1872–911

Anderson TW, Rubin H. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Stat.* 20(1):46–63

[35] Most papers use a single instrument. In the papers that use overidentified models we consider the CLR test, which is the natural alternative to AR in the multiple instrument case. We ignored papers with multiple endogenous variables, as we do not explore that case in this paper.

Andrews I, Armstrong T. 2017. Unbiased instrumental variables estimation under known first-stage sign. *Quant. Econ.* 8(2):479–503
Andrews I, Stock J, Sun L. 2019. Weak instruments in instrumental variables regression: theory and practice. *Annu. Rev. Econ.* 11:727–53
Angrist J, Kolesár M. 2024. One instrument to rule them all: the bias and coverage of just-ID IV. *J. Econom.* 240(2):105398
Angrist J, Pischke JS. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton Univ. Press
Autor DH, Dorn D. 2013. The growth of low-skill service jobs and the polarization of the US labor market. *Am. Econ. Rev.* 103(5):1553–97
Autor DH, Dorn D, Hanson G, Majlesi K. 2020. Importing political polarization? The electoral consequences of rising trade exposure. *Am. Econ. Rev.* 110(10):3139–83
Bound J, Jaeger D, Baker R. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.* 90(430):443–50
Dufour JM. 2004. Identification, weak instruments, and statistical inference in econometrics. *Can. J. Econ. Rev. Can. Econ.* 36(4):767–808
Fang H, Gavazza A. 2011. Dynamic inefficiencies in an employment-based health insurance system: theory and evidence. *Am. Econ. Rev.* 101(7):3047–77
Finlay K, Magnusson L. 2009. Implementing weak-instrument robust tests for a general class of instrumental-variables models. *Stata J.* 9(3):398–421
Finlay K, Magnusson L, Schaffer ME. 2013. *WEAKIV: Stata module to perform weak-instrument-robust tests and confidence intervals for instrumental-variable (IV) estimation*. Stat. Softw. Compon. S457684, Econ. Dep., Boston Coll., Chestnut Hill, MA
Hansen LP. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50(4):1029–54
Hansen LP, Heaton J, Yaron A. 1996. Finite-sample properties of some alternative GMM estimators. *J. Bus. Econ. Stat.* 14(3):262–80
Hornung E. 2014. Immigration and the diffusion of technology: the Huguenot diaspora in Prussia. *Am. Econ. Rev.* 104(1):84–122
Juhász R. 2018. Temporary protection and technology adoption: evidence from the Napoleonic blockade. *Am. Econ. Rev.* 108(11):3339–76
Keane M, Neal T. 2023. Instrument strength in IV estimation and inference: a guide to theory and practice. *J. Econom.* 235(2):1625–53
Keane M, Neal T. 2024. Robust inference for the Frisch labor supply elasticity. *J. Lab. Econ.* In press
Kleibergen F. 2005. Testing parameters in GMM without assuming that they are identified. *Econometrica* 73(4):1103–23
Kleibergen F, Paap R. 2006. Generalized reduced rank tests using the singular value decomposition. *J. Econom.* 133(1):97–126
Lee DS, McCrary J, Moreira MJ, Porter J. 2022. Valid *t*-ratio inference for IV. *Am. Econ. Rev.* 112(10):3260–90
Lee DS, McCrary J, Moreira MJ, Porter JR, Yap L. 2023. *What to do when you can't use '1.96' confidence intervals for IV*. NBER Work. Pap. 31893
Markevich A, Zhuravskaya E. 2018. The economic effects of the abolition of serfdom: evidence from the Russian Empire. *Am. Econ. Rev.* 108(4–5):1074–117
Moreira H, Moreira MJ. 2019. Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors. *J. Econom.* 213(2):398–433
Moreira MJ. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71(4):1027–48
Moreira MJ. 2009. Tests with correct size when instruments can be arbitrarily weak. *J. Econom.* 152(2):131–40
Nakamura E, Steinsson J. 2014. Fiscal stimulus in a monetary union: evidence from US regions. *Am. Econ. Rev.* 104(3):753–92
Olea JLM, Pflueger C. 2013. A robust test for weak instruments. *J. Bus. Econ. Stat.* 31(3):358–69
Pascali L. 2017. The wind of change: maritime technology, trade, and economic development. *Am. Econ. Rev.* 107(9):2821–54

Rao G. 2019. Familiarity does not breed contempt: generosity, discrimination, and diversity in Delhi schools. *Am. Econ. Rev.* 109(3):774–809

Shapiro JS, Walker R. 2018. Why is pollution from US manufacturing declining? The roles of environmental regulation, productivity, and trade. *Am. Econ. Rev.* 108(12):3814–54

Staiger D, Stock J. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65(3):557–86

Stock J, Watson M. 2015. *Introduction to Econometrics*. London: Pearson. 3rd ed.

Stock J, Yogo M. 2005. Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. DWK Andrews, JH Stock, pp. 80–108. New York: Cambridge Univ. Press

Voors MJ, Nillesen EEM, Verwimp P, Bulte EH, Lensink R, Soest DPV. 2012. Violent conflict and behavior: a field experiment in Burundi. *Am. Econ. Rev.* 102(2):941–64

Woolridge JM. 2012. *Introductory Econometrics: A Modern Approach*. Mason, OH: Cengage

Young A. 2022. Consistency without inference: instrumental variables in practical application. *Eur. Econ. Rev.* 147:104112