Institute for Fiscal Studies

Michael Keane
Timothy Neal

Working paper

24/46

Robust inference for the Frisch labor supply elasticity

Economic and Social Research Council

Robust Inference for the Frisch Labor Supply Elasticity

By Michael Keane and Timothy Neal*

January 26, 2024

The Frisch labor supply elasticity plays a key role in many economic policy debates, but its magnitude remains controversial. Many studies estimate the Frisch elasticity using 2SLS regressions of hours changes on wage changes. But a little appreciated power asymmetry property of 2SLS causes estimates to appear spuriously imprecise when they are shifted away from the OLS bias. This makes it difficult for a 2SLS t-test to detect a true positive Frisch elasticity. We illustrate this problem in an application to NLSY97 data. We obtain an estimate of 0.60 for young men, but the t-test indicates it is insignificant. In contrast, the Anderson-Rubin (AR) test – which is both optimal and avoids the power asymmetry problem – implies the estimate is highly significant. The same power asymmetry issue that afflicts the t-test here will arise in many IV applications. Thus, we argue the AR test should be widely adopted in lieu of the 2SLS t-test.

Keywords: *Frisch elasticity, labor supply, weak instruments, 2SLS, Anderson-Rubin test, LIML, Continuously Updated GMM*

JEL: *J22, D15, C12, C26*

I. Introduction

The Frisch elasticity measures the response of labor supply to predictable wage changes. It plays a key role in many economic policy debates because predictable wage changes have pure substitution effects. For example, Conesa, Kitao and Krueger (2009) show that a higher Frisch elasticity implies a higher optimal tax rate on capital income. And macro models where real shocks play a key role in business cycles require the elasticity to be large to match observed fluctuations in work hours over the cycle, see Prescott (2006). Because of its importance, a large literature attempts to estimate the Frisch elasticity, as in classic papers by MaCurdy (1981) and Altonji (1986) and surveyed in Keane (2011, 2021).

Classic micro data studies in the style of MaCurdy (1981) typically find the Frisch elasticity is small and insignificant,[1] while macro economists using DSGE models typically calibrate it to be in the 0.50 to 2.0 range. This has led to a long-standing "macro-micro controversy" over the magnitude of the Frisch elasticity.

We argue the classic studies were inherently biased against finding the elasticity is both large and significant, due to a combination of (i) weak instrument problems and (ii) use of biased testing procedures. The weak IV problem is well understood,

* Johns Hopkins University, CEPAR & School of Economics, UNSW.

[1]For example, MaCurdy (1981), Altonji (1986), Browning, Deaton and Irish (1985) and Ziliak and Kniesner (1999) obtain estimates for men ranging from 0.09 to 0.17.

but the testing problem arises from a little appreciated power asymmetry property of two-stage least squares (2SLS) t-tests that we explain and illustrate. Due to this power asymmetry, it is highly unlikely that any of the classic studies could have found a significant positive Frisch elasticity. Furthermore, simply using stronger instruments does not solve this problem, as it persists even when instruments are strong according to conventional weak IV testing thresholds.

The idea of the classic studies is as follows: Given panel data on workers, one may run an OLS regression of changes in log hours on changes in log wages. But this gives a downward biased estimate of the elasticity of hours with respect to *predictable* wage changes, as some changes are surprises. Instead, MaCurdy (1981) pioneered the approach of running a a 2SLS regression, using an instrument for the change in log wages with two key properties: First, it predicts wage growth at the individual level. Second, it is known at the start of the time period over which changes in wages are calculated (so it is uncorrelated with wage surprises). Then, fitted values from the first-stage of 2SLS give us predictable changes in wages, and the second stage delivers an estimate of the elasticity of hours with respect to these predictable wage changes, which is the Frisch concept.

A little appreciated property of the 2SLS estimator, first noted in Keane and Neal (2023), is that it generates a strong association between 2SLS estimates and their standard errors. This association is positive if the OLS bias is negative. Hence, positive Frisch elasticity estimates have artificially inflated standard errors. As a result, a 2SLS t-test has little power to detect a true positive Frisch elasticity, unless instruments are far stronger than is typical in this literature.

In Keane and Neal (2023) we showed that the Anderson-Rubin (AR) test largely avoids the power asymmetry that afflicts the t-test, making it a more reliable guide to 2SLS inference. The present application illustrates this point: Using NLSY97 data, we obtain a Frisch elasticity of 0.60, which is large compared to typical estimates in the classic studies. But a 2SLS t-test indicates it is not significantly different from zero. This is not surprising, as the estimate is positive and far from OLS. In contrast, the AR test indicates our estimate is highly significant (p=.018). Thus, application of a superior inferential procedure reveals clear evidence to support a fairly large Frisch elasticity value for young men.

It is well-known that the literature on estimating the Frisch elasticity using 2SLS has been hampered by weak instrument problems, as it is hard to find instruments that are strong predictors of wage growth. This issue was explored by Lee (2001), who shows how weak instrument problems in classic papers like MaCurdy (1981) and Altonji (1986) biased their estimates of the Frisch elasticity towards zero (i.e., towards OLS). Those authors used over-identified models, where the instruments used to predict wage growth were primarily age and schooling (including linear, quadratic and interaction terms). As Lee (2001) shows, given PSID samples of the size they had available, and the type of instruments they used, first-stage F-statistics in their models would have been no higher than one. This is far below conventional weak instrument testing thresholds, such as the $F > 10$ rule advocated by Staiger and Stock (1997). Thus, the instruments were very weak in the classic studies. It is well-known that with multiple weak instruments 2SLS is seriously biased towards OLS – see Bound, Jaeger and Baker (1995).

Using more recent data, Lee (2001) constructs a PSID sample five times larger than in the classic studies, and obtains a first-stage F of 26.3 and a Frisch elasticity estimate of .503 (se=.092) for 25-60 year old men.[2] In contrast, in a Monte Carlo exercise where he uses randomly drawn 1/5th sub-samples, he obtains an average elasticity estimate of only .253 (se=.227), illustrating the downward bias in the classic studies.[3] This bias towards OLS when instruments are weak is widely appreciated in the IV literature, although Lee (2001)'s result appears to have done little to alter the conventional wisdom that the Frisch elasticity is small.

We emphasize a different issue, completely unrelated to bias, that has been largely overlooked in the prior literature. In contrast to Lee (2001), we focus on the just identified case where instrument strength exceeds conventional weak IV thresholds. In particular, we show the ASVAB aptitude score in the NLSY97 is a better instrument than education, as use of this single instrument gives a first stage F above 10. In this single-strong-instrument case, 2SLS is approximately median unbiased, so bias towards OLS is not a concern (see Keane and Neal 2023; Angrist and Kolesár 2021). Nevertheless, we show that the 2SLS t-test continues to exhibit very poor behavior in this supposedly benign context.

Specifically, we show the association between 2SLS estimates and their standard errors that we document is a serious problem for inference using t-tests even when instruments are strong by conventional standards – such as the $F > 10$ criterion. The consequence is that the 2SLS t-test has little power to detect a true positive Frisch elasticity even if instruments are "strong." This will bias studies that rely on 2SLS t-tests against concluding the Frisch elasticity is large.

The implications of these results go well beyond the present application: The power asymmetry in 2SLS t-tests is a severe problem if instruments are weak, but, more importantly, it is still an important problem if instruments are quite strong. Hence, we argue the AR test should replace the t-test in 2SLS applications, not only when instruments are weak but even when they are strong.

We also present over-identified GMM results, using calculating speed and social skill measures as additional instruments for wage growth. In the over-identified case, two-step GMM suffers from the same power asymmetry problem as 2SLS. But as we also show, LIML and continuously updated GMM, in conjunction with the conditional likelihood ratio (CLR) test, resolve the problem. Our continuously updated GMM estimate of the Frisch elasticity is .55. This is only marginally significant according to the t-test, but has p=.007 according to CLR.

[2]Substantively, our estimate of 0.6 is large relative to Lee's estimate of 0.5, as our sample is much younger and there is growing evidence that the Frisch increases substantially with age - see Keane (2021). For example, French and Jones (2012) obtain an estimate of 0.36 for 40 year old men, increasing to 1.28 for 60 year olds. Imai and Keane (2004) obtain 0.36 for 25 year old men, increasing to 1.96 at age 60.

[3]Similarly, if he limits his analysis to the 1/5 of the PSID data available to the classic studies, he obtains a first stage F of only 1.06 and an elasticity estimate of .258 which is insignificant (se=.172).

Finally, to assess the generalizability of our results, we also estimate the Frisch elasticity using PSID data on men aged 20 to 65. Given the much larger PSID sample, education is a strong instrument for wage growth. We obtain an estimate of .62 using data from 1997-2019, with a CLR p-value=.000.

The outline of the paper is as follows: Section II discusses our NLSY79 data and estimating equation. Section III presents our 2SLS estimates and t-test results. Section IV presents AR test results. Section V compares the behavior of the AR and t-tests, and Section VI interprets the evidence in light of that comparison. Section VII presents results with multiple instruments, and Section VIII presents our PSID results. Section IX concludes.

II. Estimating the Frisch Labor Supply Elasticity

We estimate the Frisch elasticity using data from the National Longitudinal Survey of Youth 1997 (NLSY97). The NLSY97 follows a sample of American youth born in 1980-84. The 8,984 respondents were aged 12-17 when first interviewed in 1997.[4] We use data from rounds 10 through 15, which contain information on labor income and work hours in 2005 to 2010. The regression we run is:

$$\Delta lnH_{it} = \alpha + \beta \Delta lnW_{it} + \boldsymbol{\gamma} \mathbf{C}_{it} + \epsilon_{it} \tag{1}$$

where H_{it} is annual hours worked for respondent i in year t, W_{it} is the wage, and $\mathbf{C}_{it}$ is a vector of control variables which includes year dummies (to capture business cycle effects on hours worked) as well as respondent age and race/ethnicity.

Our hours measure is "Total annual hours worked at all civilian jobs during the year in question" while our income measure is "Annual income from wages, salary, commissions, and tips before tax deductions." We obtain an annual wage measure by taking the ratio of annual income to annual hours. Regressions that involve percentage changes can be quite sensitive to measurement error and outliers, as these can generate extreme percentage changes. So, as is typical in this literature, we implement a number of sample screens designed to eliminate outliers.[5]

Obviously, OLS estimation of (1) fails to identify the Frisch elasticity, as predictable and unpredictable wage changes have different effects on labor supply. A surprise wage increase has both substitution and income effects. In contrast, a predictable wage increase has no income effect (precisely because it was predictable), so it induces a pure substitution effect that increases labor supply. It is this Frisch substitution effect of predictable wage changes we want to estimate.

Our key task then is to choose an instrument that is known to workers at the start of each year, and that generates predictable wage growth during the year. MaCurdy (1981) and many subsequent papers use education as the primary

[4] Of that, 6748 is a random sample of the birth cohort while 2236 is an over-sample of minority groups.

[5] Observations were excluded if income was less than $\$3,000$, the annual wage was less than \$2.70 per hour worked, the total number of hours worked was less than 400 or above 4,160 (roughly 80 hours a week), or if the percentage change in wages from the last year was below -50% or above 70%.

instrument for wage growth. The motivation is that annual wage growth tends to be faster for more educated workers.[6] We adopt a closely related approach: The NLSY97 administered an aptitude test called the Armed Services Vocational Aptitude Battery (ASVAB) to respondents when they were 13 to 18 years old.[7] We find the ASVAB score is a better predictor of wage growth than education; see Online Appendix A for details. Thus, we use the ASVAB as our instrument.

We did the analysis separately for men and women, as prior literature has shown that their labor supply behavior differs in important ways. Interestingly, the ASVAB score is a much better predictor of wage growth for men than women.[8] For this reason, as well as the fact that women are more likely to be at a corner solution than men, we decided to focus only on results for men. Our full data set has 5,931 annual observations on 2,100 young men aged 22 to 30 who we observe over 2 to 6 years (the average being 3.8 years).

III. NLSY97 Estimates of the Frisch Elasticity

Table 1 presents regressions of changes in log hours on changes in log wages, as in equation (1). The first column reports OLS results. The coefficient on the log wage change is -0.42 and highly significant, with a standard error of 0.015.[9] This implies a 10% wage increase is associated with a 4.2% reduction in work hours. There are two reasons for a negative relationship: Of course surprise wage changes may generate income effects that reduce labor supply. But it is implausible that income effects alone could generate such a large negative effect.

A second key factor driving the OLS estimate negative is "denominator bias" arising because the wage is measured as the ratio of earnings to hours. If hours in the denominator are measured with error, it causes a worker's measured wage to be too low precisely when his measured hours are too high. This induces an (artificial) negative covariance between measured hours and measured wages that drives the estimated elasticity negative. As a result, the OLS estimate cannot be interpreted causally. A second virtue of instrumenting for wage changes is that it also deals with this measurement error problem - see Altonji (1986).

Next we consider the 2SLS results. The second column of Table 1 reports the first stage, where we regress log wage changes on the ASVAB percentile score to construct predictable wage changes. The coefficient is 0.039 and highly significant, with a standard error of 0.012. The effect size is substantial: A male worker in the

[6] He also used interactions of education and age, to allow the effect of education to differ by age.

[7] The ASVAB score in the NLSY97 measures mathematical and verbal aptitude. It was administered in summer 1997 to spring 1998, when the youth were aged 13 to 18 (the 13 to 14 year-olds were given an easier version). The NLS grouped respondent's into three-month age windows and calculated a youth's percentile rank within his age group.

[8] It is not clear if this is because wages grow relatively faster for high-ability men than for high-ability women – perhaps due to discrimination – or because the ASVAB is not as good a proxy for labor market skills of women. It would be interesting to explore this issue in future research.

[9] All standard errors and F-statistics reported in this paper are heteroskedasticity robust or cluster robust. The cluster robust standard errors account for both heteroskedasticity and serial correlation. They are always slightly smaller, because the errors in the hours change regression exhibit negative serial correlation. Hence the heteroskedasticty robust statistics are slightly more conservative.

Table 1—Frisch Elasticity Estimates - NLSY97

	OLS	2SLS 1^{st} Stage	2SLS 2^{nd} Stage	Reduced Form
Dependent Variable:	ΔH	ΔW	ΔH	ΔH
Wage Change	-.416		.597	
	(.015)		(.403)	
	[.015]		[.363]	
ASVAB Ability Score		.039		.024
		(.012)		(.011)
		[.011]		[.010]
F-Stat (Hetero-σ Robust)		10.12		4.47
p-value		.002		.035
F-Stat (Cluster Robust)		12.23		5.64
p-value		.001		.018
Partial R^2	.2038	.0017		.0007

Note: Heteroskedasticity robust standard errors are in parentheses. Clustered standard errors (by individual) are in square brackets. All regressions controls for year effects, age, and race/ethnicity. $N = 5,931$

100th percentile of ability is predicted to have annual wage growth 3.9 pp higher than a male worker in the 1st percentile. The heteroskdasticity robust F-test for significance of ability in the first stage regression is 10.12, which gives a p-value of 0.002. So the ASVAB instrument is significant at well above the 1% level, and passes the Staiger and Stock (1997) $F > 10$ rule of thumb for IV strength.

Notably, however, the R^2 of the first stage is only .007, implying a correlation between our predictions and actual wage changes of .084. In fact, the partial R^2 for the ASVAB test alone is .0017, implying a partial correlation of .041.[10] This illustrates the point that annual wage growth is very hard to predict. It is important to emphasize, however, that a higher first-stage R^2 would not necessarily be desirable in this context. Measured wage changes contain both unpredictable and measurement error components that we specifically want to filter out, so we expect the R^2 of the first stage regression to be far less than one.

Now consider the second stage 2SLS results, where we regress log hours changes on log predictable wage changes to obtain an estimate of the Frisch elasticity. This is reported in the third column of Table 1. The 2SLS estimate of the Frisch elasticity is 0.597, implying that a 10% predictable wage increase generates a 6% *increase* in work hours. So the use of 2SLS flips the sign of the coefficient.

[10] Recall that $F = NR^2/(1-R^2)$. Thus first-stage F depends on both the partial R^2 for the instrument and the sample size. A key insight of the weak IV literature is that asymptotic normality of the 2SLS estimator depends on F growing large rather than just N growing large.

The 2SLS estimate is clearly more reasonable: Economic theory predicts a positive Frisch elasticity, as a predictable wage increase should have a positive pure substitution effect on labor supply. And a Frisch elasticity of 0.6 is well within the range of estimates surveyed in Keane (2011, 2021), although it is clearly towards the high range of estimates for young men.

Notice however, that the (heteroskedasticity robust) standard error on the 2SLS estimate is a substantial .403, giving a t-statistic of only 1.48 and a p-value of 0.138. So, while the estimated Frisch elasticity is a substantial 0.6, it is not even significantly different from zero at the 10% level.[11] This imprecision leaves us in a quandry over what we ought to conclude from the analysis.

Should we trust the 2SLS results in this case? The first-stage F-statistic for the ASVAB instrument exceeds the commonly used weak IV threshold of 10 (if only marginally), suggesting the results may be viewed as reliable.[12]

However, weak IV tests of the type developed by Staiger and Stock (1997) and Stock and Yogo (2005) are, in the single instrument case, only designed to assess size inflation in two-tailed t-tests. Size inflation means a 5% test rejects at a rate *higher* than 5%, a positive size distortion that makes the test anti-conservative. Passing a weak IV test implies size inflation is "modest." But it does not mean the instrument is strong enough for the t-test to have acceptable power properties.[13]

In the next two sections we present AR test results, and assess the relative performance of the AR and t-tests in our data environment. We show the t-test has very poor power properties in the present application, and that the AR test ought to be relied on instead.

IV. The Anderson-Rubin Approach

Anderson and Rubin (1949) developed a different approach to inference that can also be used to test if our estimate of the Frisch elasticity is significant. The Anderson-Rubin (AR) test relies on a reduced form regression of the outcome of interest on the instrument itself, along with the control variables. In our case this is a regression of the change in log hours on the ASVAB score itself, along with the controls (time, age, race). The AR test judges the Frisch elasticity estimate to be significant if the ASVAB score is significant in the reduced form regression.

In the one instrument case, the AR test can also be obtained by running 2SLS "by hand." Run the first-stage regression of the endogenous variable x on the instrument z and the controls to obtain $\hat{x}$. Then run the second-stage regression of the outcome y on $\hat{x}$ and the controls to obtain $\hat{\beta}_{2SLS}$. The AR test of H_0:$\beta = 0$ is simply the t-test for significance of $\hat{\beta}_{2SLS}$ in the second-stage regression (t_{AR}).

[11] The cluster robust standard error is slightly smaller, at 0.363, because the serial correlation in the hours change regression is negative. But even then the t-stat is only 1.65 (p-value = 0.099).

[12] For example, Stock and Watson (2015, p.490) say: "One simple rule of thumb is that you do need not to worry about weak instruments if the first stage F-statistic exceeds 10."

[13] Stock and Yogo (2005) present critical values for F based on "worst-case" size inflation in two-tailed t-tests. For example, in the one instrument case, a sample $F > 8.96$ gives 95% confidence that a two-tailed 5% t-test will reject H_0:$\beta = 0$ at a 15% rate or less when the true β is zero. That is, it has a size inflation of no more than 10%. But passing such a test does not imply the t-test has acceptable power.

The reduced form and 2SLS "by hand" approaches to forming the AR test of $H_0{:}\beta = 0$ yield identical results. This is because regression of y on z and $\boldsymbol{C}$ and regression of y on $\hat{x}$ and $\boldsymbol{C}$ give identical t-tests for significance of z and $\hat{x}$.

The logic of the AR test is simple: A fundamental assumption of the IV method is that the instrument only affects the outcome of interest indirectly through its effect on the endogenous variable. Hence, if the instrument is significant in the reduced form, it implies that the endogenous variable has a causal impact on the outcome of interest. In our case, if the ASVAB score is significant in the reduced form, it implies that predictable wage changes influence work hours.

Of course the ASVAB score could appear significant in the reduced form merely because it somehow affects hours growth directly (not indirectly via its effect on wage growth). That is, the ASVAB score may be significant because the exclusion restriction is violated. But in that case the ASVAB score is not a valid instrument, so the 2SLS estimate and t-test results are also invalid. The very assumptions that make the IV approach valid also make the AR test valid.

The last column of Table 1 reports the reduced form results. Here, the ASVAB score is clearly significant, with a t_{AR}-stat of 2.18 (p-value 0.035). So we are left with a quandary: The AR test indicates the 2SLS estimate of the Frisch elasticity is significant, while the t-test says it isn't. Which result should we believe?

The AR test is recommended by theory as clearly superior to the t-test when instruments are weak, and no worse when instruments are strong - see Andrews, Stock and Sun (2019). This is because the AR test has three major advantages: First, it is "robust" to weak instrument problems, which means a 5% level AR test rejects a true null hypothesis at the correct 5% rate *regardless* of the strength or weakness of the instruments. In contrast, the t-test suffers from size distortions: If instruments are weak, a 5% t-test may reject a true hypothesis at rates well above or below 5%, depending on details of the situation. Second, the AR test is unbiased, meaning its power is appropriately minimized when the null corresponds to the true β. Third, Moreira (2009) shows that in the case of a single instrument (as we have here) the AR test is the most powerful unbiased test: If the null hypothesis is false, the AR test will reject the null, and conclude the parameter of interest is significant, at least as frequently as any other unbiased test.[14]

Despite its clear advantages, the AR test has been widely neglected by applied researchers. In fact, with the exception of Lee (2001), it has never been adopted in the large literature on estimating the Frisch elasticity. In the next section we present a numerical experiment based on our data that shows the performance of the AR test is *dramatically* superior in practice. We also present analytical results that lead to the same conclusion.

[14] Andrews, Stock and Sun (2019) state the advantages of the AR test more formally: "In just-identified models ... Moreira (2009) shows that the AR test is uniformly most powerful unbiased... Thus, the AR test has (weakly) higher power than any other size-α unbiased test no matter the true value of the parameters. In the strongly identified case, the AR test is asymptotically efficient ... and so does not sacrifice power relative to the conventional t-test." Moreira and Moreira (2019) extend this optimality result to models with heteroskedasticity and clustering.

V. Monte Carlo Experiment and Power Analysis

In this section we compare the AR test and the t-test to see which is a more reliable guide to the statistical significance of our Frisch elasticity estimate. To do this, we conduct the following experiment: We start from the NLS sample of N=5,931 observations that we used to generate the estimates in Table 1. We can then "bootstrap" a new artificial dataset by sampling 5,931 observations with replacement from the original sample. We do this 10,000 times to form 10,000 artifical datasets. We then repeat the analysis of Table 1, applying OLS and 2SLS to all 10,000 datasets, and summarize the results in Table 2.[15]

Table 2—Results from Monte Carlo Bootstrap Samples

	OLS		2SLS		First Stage	Reduced Form	
	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	F Statistic	$\hat{\pi}$	S.E.
Median	-0.4163	0.0146	0.5998	0.4013	10.1314	0.0238	0.0112
Mean	-0.4164	0.0146	0.7185	4.7202	11.0923	0.0237	0.0112
Std. Dev.	0.0148	0.0004	3.18636	251.4631	6.4577	0.0111	0.0003

Note: $N = 5,931$ *for each of the* $10,000$ *samples used to form the results.*

A. OLS Estimates and Standard Errors

In Table 2, both the median and mean OLS estimates of β (across all 10,000 datasets) are roughly equal to the value of -0.416 we obtained using the original NLS sample. This is expected, as our 10,000 "bootstrap" datasets mimic the covariances of the variables in the original NLS sample. The third row of Table 2 reports the standard deviation of the OLS estimates across the 10,000 artificial samples is 0.015, which equals (to three decimal places) the OLS standard error estimate reported in Table 1. Thus, the estimated OLS standard error is a very good guide to how the OLS estimates actually vary across the different samples.[16]

B. 2SLS Estimates and Standard Errors

Now we examine how the 2SLS estimates and standard errors behave. The first thing to note in Table 2 is that the median 2SLS estimate of the Frisch elasticity (across all 10,000 datasets) is 0.600, which is very close to the 2SLS estimate 0.597

[15]By sampling with replacement from the original 5,931 observations we break the panel structure of the data. As a result, the standard errors and F statistics in Table 2 will mimic the heteroskedasticity robust statistics in Table 1, not the cluster robust statistics. A natural alternative is to sample individuals rather than observations. The results of this exercise are presented in Online Appendix C and do not meaningfully change any of the conclusions presented below.

[16]Table 2 also reports the mean and median of the estimated OLS standard error across the 10,000 artificial datasets. These are again 0.015. And the variation across samples of this standard error estimate is trivially small. So the estimated standard error in each individual sample is a good guide to the actual variability of the OLS estimates across all samples.

we obtained using the original NLS dataset. This is exactly as expected: As our artificial datasets are constructed from our original NLS sample, we can think of the NLS sample as the "population" from which all 10,000 datasets are drawn. In this population, 0.597 is in fact the true value of the Frisch elasticity. We see that the median 2SLS estimate accurately uncovers the true Frisch elasticity value. As Keane and Neal (2023) and Angrist and Kolesár (2021) show, the median bias of 2SLS is negligible at this level of instrument strength.

Second, note that the median of the estimated 2SLS standard errors, reported in the first row of Table 2, is 0.401. This agrees closely with the 2SLS standard error estimate of 0.403 in Table 1. However, the actual empirical standard deviation of the 2SLS estimates across the 10,000 data sets is 3.864. In contrast to OLS, the actual variation of the 2SLS estimates bears no resemblance to the estimated 2SLS standard errors. This is our first indication that the 2SLS standard errors are not a good guide to the actual variability of the 2SLS estimates across samples.[17] This in turn means that 2SLS t-statistics – which rely on those standard error estimates – will not be a useful guide to significance of 2SLS estimates.

To further explore the behavior of the 2SLS standard error, Figure 1 plots the 2SLS standard errors against the 2SLS estimates of the Frisch elasticity from each of the 10,000 samples. A striking aspect of the figure is the strong positive association between the 2SLS estimates and their standard errors: The Spearman correlation is an extraordinarily large 0.905. This means that in samples where the estimated Frisch elasticity is larger, the standard error is also larger.[18] As we will see, this pattern has extremely important empirical implications.

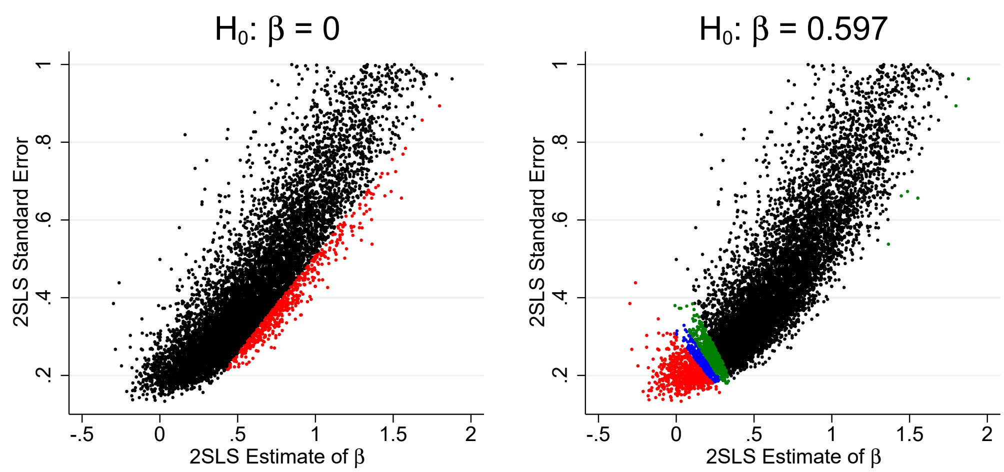
The association between 2SLS estimates and their standard errors is not specific to this application. It is a generic but little-appreciated property of the 2SLS estimator. Exact finite sample theory sheds some light on this phenomenon. Phillips (1989) shows that in the unidentified case the 2SLS variance estimator $\hat{\sigma}^2 = (N-2)^{-1} \sum_i (y_i - x_i \hat{\beta}_{2SLS})^2$ converges in distribution to a quadratic function of $\hat{\beta}_{2SLS}$, with a minimum at $E(\hat{\beta}_{OLS})$. Thus, the standard error of regression ($\hat{\sigma}$) is minimized when $\hat{\beta}_{2SLS}$ is close to $E(\hat{\beta}_{OLS})$. Of course, the standard error of the regression ($\hat{\sigma}$) is a fundamental driver of the standard error of $\hat{\beta}_{2SLS}$, which is $\hat{\sigma}/\sqrt{N \hat{cov}(z,x)^2/\hat{\sigma}_z^2}$. Thus, in the unidentified case, the standard error of $\hat{\beta}_{2SLS}$ tends to be minimized when the estimate is near $E(\hat{\beta}_{OLS})$.

Importantly, this property of 2SLS in the unidentified case still influences the behavior of 2SLS estimates and standard errors in strongly identified models. In fact, Phillips (1989) calls this the "leading case" as it provides the leading term of the series expansion of the density of the estimator in the general case. As a result, even in strongly identified models, the standard error of $\hat{\beta}_{2SLS}$ tends to be minimized when the estimate is near $E(\hat{\beta}_{OLS})$, as we see in Figure 1. We

[17] Of course, in the single instrument case the mean and variance of the 2SLS estimator do not exist, which means that if we did many more than 10,000 runs the mean and variance wouldn't converge. Hence, the standard deviation of the 2SLS standard error cannot be bootstrapped.

[18] A graph of OLS standard errors vs. estimates is a spherical cloud, as they have zero correlation.

FIGURE 1. STANDARD ERROR OF $\hat{\beta}_{2SLS}$ PLOTTED AGAINST $\hat{\beta}_{2SLS}$ ITSELF



Note: Runs with standard error > 1 are not shown. In the left panel, red dots indicate $H_0{:}\beta = 0$ is rejected at the 5% level by a 2SLS t-test. In the right panel red dots indicate $H_0{:}\beta = 0.597$ is rejected at the 5% level. Blue and green indicate 10% and 20%.

illustrate this phenomenon via simulations in Online Appendix D.

A second factor is also at play: The 2SLS standard error also depends on $\hat{cov}(z, x)$, the sample covariance of the instrument and the endogenous variable. In the population $cov(z, x) = 0$. But in finite samples where the exogenous instrument happens (by chance) to be *positively* correlated with the structural error in the outcome equation, the 2SLS estimate is shifted toward OLS and the instrument appears stronger, so the estimate seems more precise.[19] Appendix A.4 provides additional mathematical detail on why this pattern arises. For our present purposes, it suffices to note the following: Because of this pattern, large positive 2SLS estimates of the Frisch elasticity will have relatively large standard errors, while estimates near zero will have small standard errors.

[19] A more detailed intuitive explanation for the association between 2SLS estimates and their standard errors is as follows: As we discussed in Section III, in the original NLS sample the partial correlation between the ASVAB score and wage growth is 0.04. But the correlation fluctuates across our 10,000 subsamples due to sampling variation. Two things happen in samples where it is relatively high:

First, the 2SLS standard error estimate is smaller: The higher the correlation between the instrument and the endogenous variable, the stronger the instrument seems to be, and hence the smaller is the 2SLS standard error. 2SLS standard errors are suspect due to this pattern.

Second, the 2SLS estimate is more shifted in the direction of the OLS bias (which is negative). This is because, as we discussed in Section III, if the predictable part of wage growth is small, then a high correlation between the instrument and the endogenous variable is not really a good thing. In samples where that correlation rises above 0.04, the instrument is picking up some of the endogenous part of wage growth that arises due to measurement error and surprise wage growth. This in turn means the 2SLS estimate will be shifted in the direction of the OLS bias (negative).

Putting these two facts together, we get that 2SLS estimates that are most shifted in the direction of the OLS bias (negative) appear to be more precise. This is exactly the pattern we see in Figure 1.

This bring us to our key point: The positive association bewteen 2SLS estimates of the Frisch elasticity and their standard errors has important implications for statistical inference. As we now show, this mechanical relationship makes it very difficult for a 2SLS t-test to detect a true positive Frisch elasticity.

Recall that our 10,000 simulated data sets are constructed so the true value of the Frisch elasticity in these data sets is 0.597. Thus, if the 2SLS t-test is reliable it should have two properties: First, if we run 5% t-tests of the hypothesis that the true Frisch elasticity is zero we should reject that false hypothesis at a high rate (indicating the test has good power). Second, if we run 5% t-tests of the true hypothesis that the Frisch elasticity is equal to 0.597 (the true value) we should reject that hypothesis approximately 5% of the time (indicating the test has correct size). Furthermore, those rejections should be evenly split between cases where the estimated Frisch elasticity is above and below the true value.

In the left panel of Figure 1 we shade in red the cases where the 2SLS t-test rejects the false null hypothesis that the true Frisch elasticity is equal to zero. These are the cases where the ratio of the estimate to the standard error exceeds the 5% critical level of 1.96 (in absolute value). Notice that the red shaded area is quite small. In fact, the false null hypothesis is only rejected 5.1% of the time. This is an extremely low level of power. It scarcely the exceeds the 5% rate at which a well-behaved 5% level test should reject a *true* null hypothesis.

In the right panel of Figure 1 we shade in red the cases where the 2SLS t-test rejects the true null hypothesis that the Frisch elasticity equals its true value of 0.597. The test rejects the null hypothesis 6.6% of the time. This is not bad when viewed in isolation, as it is not too far from the correct rate of 5%, so the t-test size distortion is small. But more importantly, the rate of rejecting the true hypothesis that the Frisch elasticity equals 0.597 is actually *greater* than the rate of rejecting the false hypothesis that the elasticity equals 0. This is extremely poor behavior for a statistical test: The fact that size exceeds power means the t-test is uninformative about the true parameter value.

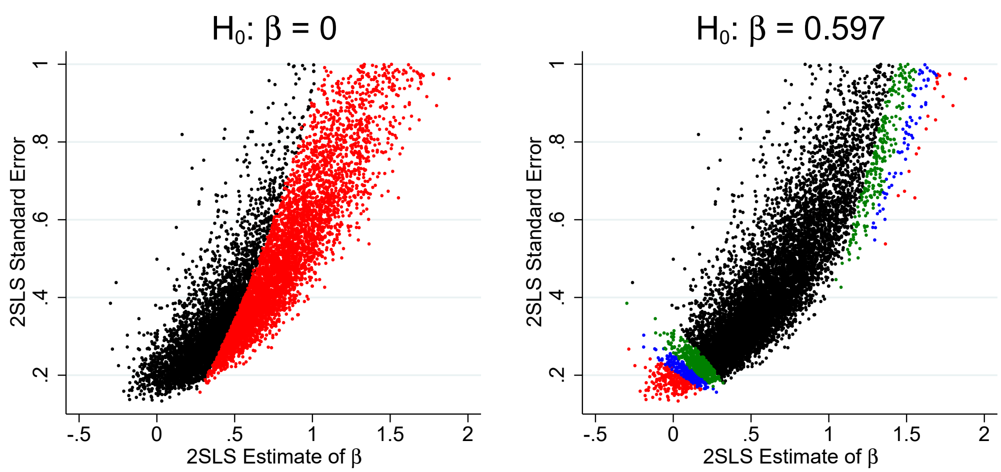
These results illustrate our key point: The true F easily passes conventional weak instrument testing thresholds of the type developed by Staiger and Stock (1997) and Stock and Yogo (2005). So, as expected, the median bias of 2SLS and the size distortion in the t-test are trivial. Nevertheless, the 2SLS t-test results are completely uninformative, as size exceeds power.

Another notable aspect of the right panel of Figure 1 is that the cases where we reject the null of $\beta = 0.597$ are not evenly split between cases where the estimate is above and below the true value. In fact, all the rejections occur when the estimated Frisch elasticity is very small (near zero). This is a direct consequence of the positive association between 2SLS estimates and their standard errors. As large positive estimates of the Frisch elasticity have large standard errors, there is very little chance of concluding a large positive estimate is significant. We refer to the low power of the 2SLS t-test to detect a true effect opposite in sign to the OLS bias as the "power asymmetry" problem.

C. *Anderson-Rubin Test Results*

Figure 2 reports the same results for the AR test. The contrast with the t-test is dramatic. We again plot the 2SLS standard errors against the 2SLS estimates, as in Figure 1. But now we plot in red the cases where the AR test (t_{AR}) rejects the false null hypothesis that $\beta = 0$ at the 5% level. In the left panel we see that the red region is quite large. The AR test rejects the false null that the Frisch elasticity is equal to zero 56.5% of the time. This is a good level of power that is more than ten times greater than the 5.1% rate achieved by the t-test.

FIGURE 2. STANDARD ERROR OF $\hat{\beta}_{2SLS}$ PLOTTED AGAINST $\hat{\beta}_{2SLS}$ ITSELF (AR TEST)



Note: Runs with standard error > 1 are not shown. In the left panel, red dots indicate H_0:$\beta = 0$ is rejected at the 5% level by the AR test. In the right panel red dots indicate H_0:$\beta = 0.597$ rejected at the 5% level. Blue and green indicate 10% and 20%.

The right panel of Figure 2 shows how often the AR test rejects the *true* null hypothesis that the elasticity equals 0.597. The 5% level AR test rejects 4.9% of the time, which is almost exactly equal to the correct 5% rate. Furthermore, we plot in blue and green the cases where 10% and 20% AR tests reject. These rates are 10% and 19.5%, so again almost perfect. This illustrates how the AR test is "robust" in the sense that it has correct size (rejection rates) regardless of the strength or weakness of the instruments. Thus we see that *the AR test has correct size and ten times the power of the t-test.*

The only limitation of AR is it doesn't quite generate symmetric rejections when the estimates are above and below the true value. For example, of the 4.9% rejections by the 5% test, 3.6% occur when the estimate is below 0.597 and 1.3% occur when it is above. Like the t-test, the AR test tends to attribute greater precision to estimates shifted in the direction of the OLS bias, and less precision to large positive estimates. But Keane and Neal (2023) show the AR power asymmetry vanishes quickly as instruments become stronger, while the t-test power asymmetry remains substantial even with very strong instruments.

These results make it very obvious that in the data environment of our empirical application the AR test provides a far more reliable guide to the significance of the estimate of the Frisch elasticity than does the t-test. The consequence is that prior work that relied on 2SLS t-tests will have tended to obtain insignificant results even if the true Frisch elasticity is well above zero.

The superiority of the AR test over the t-test is not specific to this example. Due to the power asymmetry that plagues the t-test, it has, in general, difficulty detecting true negative (positive) effects when the OLS endogeneity bias is positive (negative). And this problem is relevant even when instruments are much stronger than here. The AR test is much less susceptible to this problem.

D. *Analytical Power Function Comparison: AR vs t-test*

To show the generality of the problem we have described, we now compare the analytical power functions of the AR and t-tests in an exactly identified linear IV model with *iid* normal errors:

$$\begin{gathered} y = \beta x + u \\ x = \pi z + e \quad \text{where} \quad e = \rho u + \sqrt{1-\rho^2}\eta \end{gathered} \tag{2}$$

where $u \sim iidN(0,1), \eta \sim iidN(0,1)$, and $z \sim iidN(0,1)$. Thus the instrument z satisfies $cov(z,u) = 0$ and $cov(z,\eta) = 0$. The parameter $\rho \in (-1,1)$ determines the severity of the endogeneity problem, while π determines the strength of the instrument.[20] We normalize $Var(e) = 1$ so β can be interpreted as roughly the standard deviation change in y induced by a 1 standard deviation change in x.

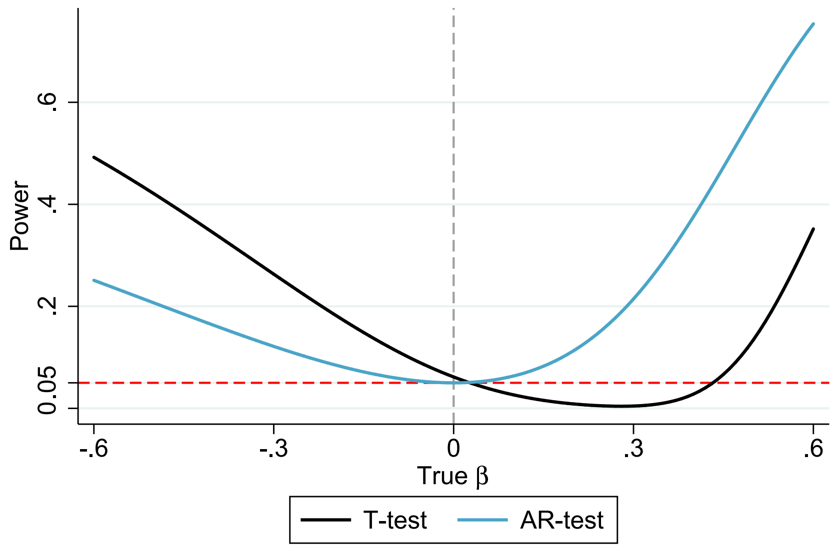
Figure 3 compares the power functions of the AR and t-tests, obtained via the procedure described in the Appendix, with true F=10.12 and $\rho = -.70$ to mimic our empirical application. The power function is the probability a 5% level test rejects H_0:$\beta = 0$, conditional on each alternative true β listed on the x-axis.

Several features of the power functions are notable. First, power of the 2SLS 5% level t-test to reject H_0:$\beta = 0$ is close to 5% if the true $\beta = 0$. So the t-test has roughly correct size if F=10 and ρ=-.7. Both Angrist and Kolesár (2021) and Keane and Neal (2023) argue size inflation in two-tailed t-tests is modest unless instruments are very weak and endogeneity is very severe, and that is reflected here.[21] If size inflation were the only concern, it would be fine use the t-test.

[20] This *iid* normal setup is not as restrictive as it may appear, as Andrews, Stock and Sun (2019) show that for any heteroskedastic DGP, there exists a homoskedastic DGP yielding equivalent behavior of 2SLS estimates and test statistics. Any exogenous covariates can be partialed out of y and x without changing anything of substance. And the variance normalization is without loss of generality as one can always normalize y, x and z to have variance of one.

[21] Stock and Yogo (2005) derived critical values that the sample F must surpass in order for a researcher to have high confidence the "worst-case" size inflation in the two-tailed t-test is modest. But the "worst case" occurs when ρ is near one or minus one, so endogeneity is extremely severe. For smaller values of ρ – that are more typical of applications – much lower levels of F suffice to render size inflation modest. Keane and Neal (2023, 2022) and Angrist and Kolesár (2021) both make this point.

FIGURE 3. POWER OF THE T-TEST VS. AR-TEST WHEN TRUE F = 10 ($\rho = -0.7$)



Note: Probability a 5% level test rejects $H_0 : \beta = 0$, conditional on each alternative true β listed on the x-axis. In order to most closely match our application, we set $\rho = -0.7$ and the true first-stage F-statistic to 10.12.

Second, however, the severe bias of the t-test is evident: It's power dips below 5% for a wide range of positive values of β, dipping to near zero when true β is 0.20 to 0.40. In the model in (2), β is approximately the standard deviation change in y induced by a one standard deviation change in x. So effect sizes of 0.20 to 0.40 would be substantial in most applications. Yet the 2SLS t-test has essentially no power to detect effect sizes in this range. In contrast, the unbiasedness of the AR test is evident in Figure 3, as its power is appropriately minimized at $\beta = 0$.

Fourth, and most importantly, the AR test has far superior power to detect true positive values of β in this environment. The lack of power of the t-test in the positive β range is due to the positive association between 2SLS estimates and their standard errors that exists when the OLS bias is negative. This causes larger positive estimates of β to have spuriously inflated standard errors.

In Appendix A.2 we show the power curve of the AR test is *identical* to that of the t-test from the infeasible IV regression of y on πz that one could run if π were known. This makes clear why AR is optimal, and superior to the t-test that relies on an estimate of π. In Figure 3 the t-test *appears* to have better power than AR when the true β is negative, but this is deceptive: it occurs because negative estimates of β tend to have spuriously small standard errors.

Appendix A.6 presents additional results at different levels of F. We show the power asymmetry is more pronounced when instruments are weak, but persists even when instruments are far above conventional weak IV testing thresholds.

VI. Interpreting the Empirical Results in Light of the Experiment

We now return to the empirical results in Table 1, and reassess them based on what we have learned from both the Monte Carlo and analytical results presented in Section V. Recall our 2SLS estimate of the Frisch elasticity is 0.597, which is well above those obtained in the classic studies.[22] But the 2SLS t-test indicates it is not significantly different from zero at the 5% level (p=.14 and p=.10 for the heteroskedasticity and cluster robust t-tests). However, the analysis of Section V shows that the t-test has little power to detect true positive effects of plausible magnitude in this data environment, that is characterized by (i) a first stage F slightly above 10 and (ii) a correlation between the reduced form residuals of -0.70, so that the OLS bias is strongly negative.

The analysis of Section V revealed that the AR test is far more reliable that the t-test in our context. It indicates that our Frisch estimate of 0.597 is significant at the 3.5% or 1.8% level, depending on whether we rely on the heteroskedasticity or cluster robust version of the test. We can also invert the AR test to obtain a weak instrument robust confidence interval, as discussed in Anderson and Rubin (1949).[23] Using cluster robust statistics we obtain a 95% confidence interval for the Frisch elasticity of 0.082 to 2.03, which is clearly bounded above zero, and covers most of the range often used to calibrate macro models.

A. *One-Tailed Test Results*

Theory suggests the Frisch elasticity must be positive. So it makes sense to consider one-tailed tests of $H_0{:}\beta \leq 0$. Given our sample size and degrees of freedom, a 5% level one-tailed t-test rejects $H_0{:}\beta \leq 0$ if $t > 1.645$.

In the one instrument case recall that the AR test is the t-test for significance of β_{2SLS} in the OLS regression of y on $\hat{x}$ and the control variables – what one obtains by running the 2nd stage of 2SLS "by hand." We call this t_{AR}. A 5% level one-tailed AR test rejects $H_0{:}\beta \leq 0$ if $t_{AR} > 1.645$.

Table 3—Frisch Elasticity: One-Tailed Tests

	t-stat	p-value	t_{AR}-stat	p-value
Hetero-σ Robust	1.482	0.069	2.113	0.017
Cluster Robust	1.648	0.050	2.374	0.009

Note: Based on regressions presented in Table 1.

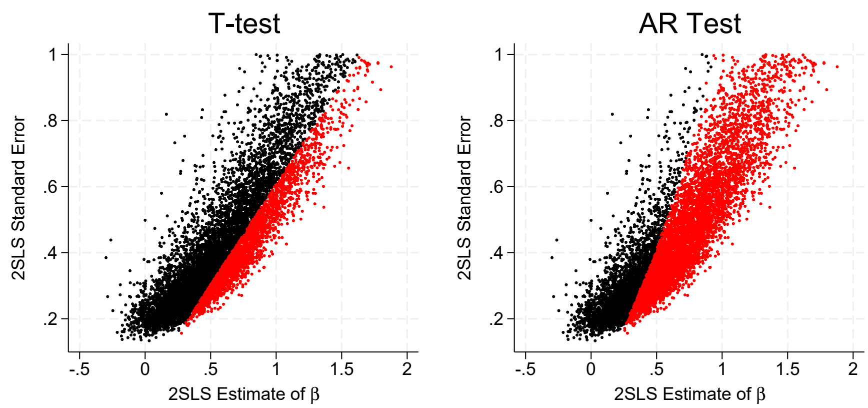
One-tailed test results are presented in Table 3. Using heteroskedasticity robust standard errors, the 2SLS t-stat (1.48) indicates significance at the 6.9% level, while the AR test (t_{AR}) indicates significance at the 1.7% level.

[22] In Online Appendix B we show it is robust to alternative sample screens and sampling weights.

[23] The basic idea of AR test inversion is to run regressions of $y - xb$ on the instrument and control variables, and find the lower and upper cutoffs for b where the AR test p-value is exactly .05.

Figure 4 repeats the Monte Carlo experiment from Section V of the main text. Given a true β of .597, the t-test rejects the false null hypothesis H_0:$\beta \leq 0$ only 23.5% of the time, while the AR test rejects at a 68.8% rate. Thus, the AR test has 3 times the power of the t-test in this environment.

FIGURE 4. STANDARD ERROR OF $\hat{\beta}_{2SLS}$ PLOTTED AGAINST $\hat{\beta}_{2SLS}$ ITSELF



Note: Runs with standard error > 1 are not shown. In the left (right) panel, red dots indicate H_0:$\beta \leq 0$ is rejected at the 5% level by a one-tailed t-test (AR Test).

Figure 5 presents analytic power curves for the one-tailed AR and t tests with true F=10.12 and $\rho = -.70$ to mimic our empirical application. The power function is the probability a 5% level test rejects H_0:$\beta \leq 0$, conditional on each alternative true β listed on the x-axis. When β is in the 0 to 0.30 range a one-tailed 2SLS t-test has essentially zero power, due to the power asymmetry. And its power remains well below the AR test at higher levels of true β.

FIGURE 5. POWER OF ONE-TAIL t-TEST VS. AR-TEST, TRUE F $= 10$ ($\rho = -0.7$)

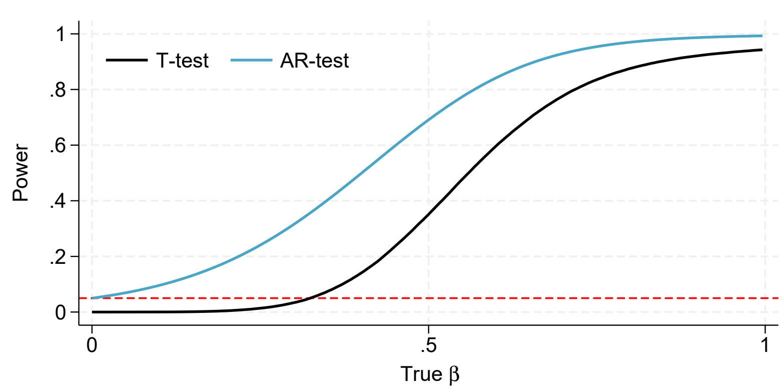
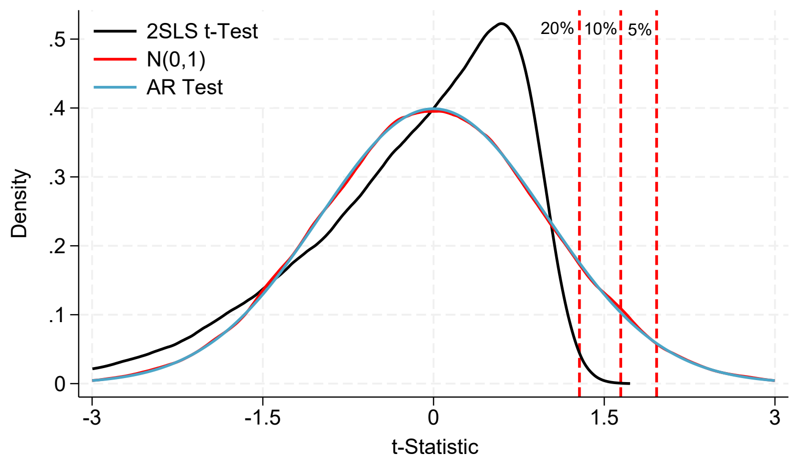


Figure 6 gives further insight into the poor performance of the 2SLS t-test by comparing its density to the AR test. Given a true $\beta = 0$ the distribution of t_{AR} is almost exactly a $N(0,1)$ random variable. This is of course as expected for a t-test with roughly 1000 degrees of freedom.

In contrast, the 2SLS t-test is highly non-normal: It has little mass above the 20% critical value, and literally none above the 10% value. And it is highly skewed left, the direction of the OLS bias. It could be called an "ersatz t-test" as it does not remotely approximate a normal distribution.

FIGURE 6. DENSITY OF t-STAT VS. AR TEST, TRUE F = 10, $\rho = -0.7$, TRUE $\beta = 0$



B. Extensions

Going beyond our particular estimation result, our broader point is that the power problems that afflict the 2SLS t-test make it difficult to detect a true positive Frisch elasticity. Thus, we argue it is important that future work on estimating the Frisch elasticity should rely on the AR test rather than the t-test.

Our results here also provide lessons that are useful in a broader context. In general, 2SLS t-tests have low power to detect true effects that are opposite in sign to the OLS bias, even if instruments are very strong by conventional standards – see Appendix A for more details. For this reason we argue the AR test should replace the t-test for inference in exactly-identified linear IV models, not only when instruments are weak but even when they are strong.[24] Next we consider over-identified models, where other alternatives to the t-test are available.

[24]The AR test is widely recommended by theorists for use when instruments are weak, because it is robust to weak instrument problems. That is, the AR test is guaranteed to have correct size when instruments are weak, yet it is guaranteed to be no less powerful than the t-test when instruments are strong. But our analysis in Sections V and has revealed that the AR test has greatly superior power properties to the t-test even when instruments are strong by conventional standards (i.e., F >10).

VII. Results Based on Multiple Instruments

In this section, to gain efficiency, we report results from over-identified models where, in addition to the ASVAB score, we use two additional instruments:

One is a numerical speed test that is included in the ASVAB test battery, but that the NLS excludes when constructing their summary ASVAB percentile score. The speed test assesses how many simple mathematical computations a respondent can answer correctly in a short period of time. It has only a modest correlation with the ASVAB score, and it independently predicts wage growth, suggesting it captures a different dimension of cognitive ability.

In addition, as a measure of non-cognitive skill, we also include the NLSY97 delinquency index. Like the ASVAB score it is time invariant and measured once in 1997. It is an index from 0-10 which measures the number of anti-social or criminal behaviors the respondent has done at any point in their lives up until that point. Examples include joining a gang, stealing, carrying a handgun, selling drugs, or participating in a serious fight.[25]

As we see in the first column of Table 4, the speed test is actually more significant than the ASVAB score in the first stage of 2SLS.[26] The delinquency score has the expected negative sign, but it is not individually significant.

We also report the heteroskedsaticity and cluster robust partial F-statistics for joint significance of the instruments in the first stage, as well as the Olea-Pflueger effective F-test for weak instruments in a non-iid setting. These statistics range from 5.3 to 6.9, so they are well below conventional weak instrument testing thresholds.[27] Thus weak instruments are clearly a concern and the 2SLS t-test cannot be viewed as reliable.

The 2SLS estimate of the Frisch elasticity is .468, which is smaller than the estimate of 0.597 we obtained in Table 1. This is consistent with the idea that use of multiple instruments tends to generate a bias towards OLS. Notably, the heteroskedasticity robust standard error decreases from 0.403 to 0.296, reflecting the efficiency gain from adding additional relevant instruments in the first-stage. But t=1.58 (p=.11), so a 5% t-test judges our estimate insignificant.

In Keane and Neal (2023) we show the power asymmetry problem that plagues the t-test is more severe in the over-identified case, as are size distortions. Thus it becomes even more important to use robust tests. So now we consider the AR test, which in the over-identified case is simply the F-test for joint significance of the three instruments in the reduced form. As we see in Table 4 the cluster robust AR test is 3.23, with a p value of .022.

[25]We also tried other measures of non-cognitive skill, such as the NLSY97 mental health index, but found that the delinquency score had the strongest correlation with subsequent wage growth.

[26]In order to keep the sample identical to that in Table 1, we set the speed test and delinquency scores to their mean in the few cases where they are missing.

[27]As Andrews, Stock and Sun (2019) note, in over-identified models it is inappropriate to use a heteroskedasticity-robust or conventional F-test to assess instrument strength in non-homoskedastic settings. They suggest the Olea and Pflueger (2013) effective first-stage F-statistic. In the single instrument case we considered in Sections III-24, this reduces to the conventional heteroskedasticity-robust F.

Table 4—Frisch Elasticity - Over-identified Models

	2SLS 1^{st} Stage	2SLS 2^{nd} Stage	Reduced Form	GMM-2S 2^{nd} Stage
Dependent Variable	ΔW	ΔH	ΔH	ΔH
Wage Change		.468		.469
		(.296)		(.296)
		[.261]		[.261]
ASVAB Ability Score	.020		.023	
	(.015)		(.014)	
	[.014]		[.012]	
Numerical Speed Test	.014		-.002	
	(.007)		(.006)	
	[.006]		[.006]	
Delinquency Score	-.002		-.003	
	(.002)		(.002)	
	[.002]		[.002]	
F-Stat (Hetero-σ Robust)	5.48		2.39	
p-value	.001		.066	
F-Stat (Cluster Robust)	6.92		3.23	
p-value	.000		.022	
Olea-Pflueger Effective F	5.26			
Exogeneity Test (Sargan or J)		1.65		1.63
p-value		.439		.442
Partial R^2	.0028		.0012	

Note: 'GMM-2S' refers to 2-step GMM. Heteroskedasticity robust standard errors are in parentheses. Clustered standard errors are in square brackets. All regressions control for year effects, age, and race/ethnicity. $N = 5,931$.

Thus we see a pattern similar to Table 1: The t-test implies the Frisch elasticity estimate is not significant at the 5% level, while the AR tests implies a much higher level of confidence.[28] The weakness of the 2SLS t-test result is again attributable to the positive covariance between 2SLS estimates and standard errors, which makes it difficult for 2SLS t-tests to detect a positive Frisch elasticity.

We can obtain a Sargan test of the 2SLS over-identifying restrictions by regressing the 2SLS residuals on the full set of instruments and exogenous variables. The NR^2 of the regression is distributed $\chi^2(K)$ under the null $\beta = 0$, where K is the number of over-identifying restrictions. In Table 4, we see the Sargan test statistic is 1.65. It is distributed $\chi^2(2)$ so the p-value is .439. Thus we cannot reject

[28]If we invert the AR test (cluster robust F version) we obtain a 95% confidence interval for the Frisch elasticity of 0.061 to 1.962. This interval sits above zero, and covers the range of values typically used to calibrate macro models.

the exogeneity of the instruments. This is important, as a failure of the over-identification test would invalidate the AR test, as it would indicate that the instruments may be significant in the reduced form merely because they affect hours changes directly (rather than only indirectly via wages as 2SLS assumes). So the AR and Sargan statistics should be evaluated in conjunction. Of course, failure of the exogeneity test would invalidate 2SLS t-test results as well, so this is not a disadvantage of AR relative to the t-test.

The last column of Table 4 reports the two-step GMM results. Two-step GMM and 2SLS are equivalent under homoskedasticity, but GMM increases efficiency in the heteroskedastic case. The two-step GMM estimate of 0.469 is virtually identical to the 2SLS estimate, suggesting that heteroskedasticity has a minimal impact on the results. The Hansen-J test has a p-value of 0.44, so again we cannot reject exogeneity of the instruments.

While it is standard to present results like those in Table 4 in the over-identified case, we argue they should not be relied on, for two key reasons:

First, as we already noted, it is well known that both 2SLS and GMM-2S are biased towards OLS in the over-identified case.[29] Fortunately, the limited information maximum likelihood (LIML) estimator of Anderson and Rubin (1949) does not suffer from this bias problem. And the continuously updated GMM estimator of Hansen, Heaton and Yaron (1996), that we denote GMM-CU, generalizes LIML to heteroskedastic data. LIML and GMM-CU are identical under homoskedasticity, so the reason for reporting GMM-CU is to gain more efficient estimates in the presence of heteroskedasticity and clustering.

The theory literature recommends using LIML with weak instruments, but in Keane and Neal (2023) we show that LIML performs much better than 2SLS even when when instruments are strong by conventional standards (e.g., when thresholds like $F >10$ are satisfied).

Second, in the over-identified case, the AR test is no longer optimal, so choosing the right test is more complicated. Under homoskedasticity, the choice is clear, as the weak instrument robust conditional likelihood ratio (CLR) test of Moreira (2003) is optimal.[30] We refer to this as CLR-LIML. Kleibergen (2005) extends the CLR test to GMM, making it robust and more efficient given heteroskedasticity and/or within-cluster correlation. We refer to this as CLR-GMM. See Online Appendix E for further discussion of tests in the over-identified case.

[29]Lee (2001) argues that this may have caused the classic studies, which used over-identified models, to obtain estimates of the Frisch elasticity near zero.

[30]We explain the CLR test in detail in Keane and Neal (2023). It is based on the reduced form system in which the two endogenous variables, hours changes and wage changes, are regressed on all the exogenous variables. The instrument exclusion condition implies the coefficients on the excluded instruments in the hours equation are β times those in the wage equation. So this proportionality constraint is imposed when estimating the reduced form system. A likelihood ratio (LR) test assesses the deterioration in the log-likelihood of the system when the additional constraint $\beta = 0$ is imposed. The likelihood here takes the simple form of a sum of squared residuals, so one is actually assessing how much the residual variance increases. The test is $\chi^2(1)$. It was originally proposed by Anderson and Rubin (1949). Moreira (2003) showed how to adjust the critical values of this LR test so it is robust to weak instruments, giving the CLR test. A key point is that in the single instrument case AR and CLR are equivalent.

In Online Appendix F we present Monte Carlo experiments to assess the power of the various tests available in the over-identified case. These experiments use bootstrapped artificial samples from the NLSY97, just like the experiments in Section V and Online Appendix C.[31] If we sample by observation as in Section V, we find that a 5% 2SLS t-test rejects the false null H_0:β=0 at an 12.9% rate, compared to 40.6% for the AR test, and 54.3% for CLR-GMM. So the ranking is as expected, with both AR and CLR exhibiting much better power than the t-test.[32] The power of CLR-GMM is best, consistent with our recommendation.

Importantly, the Monte Carlo shows that two-step GMM suffers from the same power asymmetry as 2SLS, due to positive correlation between GMM-2S estimates and their standard errors. The GMM t-test is therefore unreliable as well. Thus, we recommend that applied researchers rely on GMM-CU and CLR-GMM in the over-identified case if heteroskedasticity/clustering are concerns.

Table 5—Frisch Elasticity - Robust Estimators and Tests

	LIML	GMM-CU
Wage Change	.567	.552
	(.349)	(.311)
	[.309]	[.279]
CLR-LIML Test	5.55	
p-value	.022	
CLR-GMM Test		7.69
p-value		.007
S Statistic		10.07
p-value		.018
Exogeneity Test (Sargan or J)	1.45	1.55
p-value	.484	.460

Note: 'GMM-CU' refers to continuously updated GMM. Heteroskedasticity robust standard errors are in parentheses and clustered standard errors are in square brackets. All regressions controls for year effects, age, and race/ethnicity. $N = 5,931$.

Table 5 presents LIML and GMM-CU results. The LIML and GMM-CU estimates of the Frisch elasticity are 0.567 and .552, respectively. As expected, these are larger than the 2SLS and GMM-2S estimates of 0.468 and .469, given that the

[31]In the one instrument case the instrument is uncorrelated with the 2SLS residuals. So when we treat the full sample as the "population," the instrument has zero population covariance with the structural error by construction. But in the over-identified case the instruments do have small correlations with the 2SLS residuals. We need to partial out those correlations to set up the experiment.

[32]If we sample individuals instead of observations in the experiment, and use cluster robust standard errors, we obtain 29.0% for t, 51.7% for AR and 67.2% for CLR-GMM. Power is higher due to negative autocorrelation in the errors, which is also why cluster-robust standard errors are smaller in Table 5.

later are biased towards OLS. The LIML standard error is substantial, implying that the LIML estimate is at best marginally significant. But the LIML standard error is no more reliable than the 2SLS standard error, because, as we explain in Keane and Neal (2023), the LIML estimates and standard errors have the same positive association that afflicts 2SLS estimates and standard errors.

The CLR-LIML test statistic is 5.55, and, as it has a $\chi^2(1)$ distribution, the p-value is 0.022. The CLR-GMM statistic is 7.69 with p-value of 0.007. Thus the evidence for a positive Frisch elasticity based on the robust tests is strong. The CLR-GMM 95% confidence interval is 0.133 to 1.583, so it sits comfortably above zero but covers a wide range.

Stock and Wright (2000) develop an alternative robust test that generalizes the AR test to the GMM case. This "S-statistic" is the GMM objective function evaluated at $\hat{\beta}$=0. For GMM-CU we find S=10.07. The test is distributed $\chi^2(3)$ so the p-value is .018. Thus the Frisch elasticity estimate is again significant.

Finally, we consider Hansen's test of over-identifying restrictions. As we see in Table 5 the J-test has p >0.40, indicating we cannot reject the exogeneity of the instruments. This is important, as a failure of the J-test would invalidate the tests and estimators in this table.[33]

In summary, in the over-identified case our preferred GMM-CU estimate of the Frisch elasticity, .552, is very similar to the .597 estimate we obtained using 2SLS with a single instrument. The weak instrument robust AR, CLR-GMM and S tests all indicate the GMM-CU estimate is highly significant. The use of three instruments does increase efficiency, as the confidence interval narrows from the AR interval of (.082, 2.03) we reported in Section 24 to (.133, 1.583) here.

VIII. PSID Results

Finally, we present results using the PSID. This enables us to check if our estimate of the Frisch elasticity generalizes to a different panel data set that is much larger and includes workers of all ages. We now use years of education as the instrument for wage growth, as no ability measure like the ASVAB is available.

Unfortunately, the PSID was annual from 1968 to 1997 but biennial afterwards. To deal with this we consider both a biennial sample from 1997 to 2019, and a longer sample from 1975 to 2019, dropping every second year so the whole analysis sample is biennial. Thus, we look at effects of two-year wage changes on two-year hour changes. We refer to these as the recent and long samples.

The recent sample contains 34,670 observations on 7,938 men aged 20 to 65, an average of 4.4 biennial observations each. The long sample contains 64,701 observations on 12,559 men, an average of 5.2 biennial observations each. As we see in Table 6, the cluster robust first-stage F statistic for the education

[33] In the single endogenous variable, K instrument case, the Sargan and J-tests have power to detect if at least one instrument is endogenous, provided the model is over-identified, which means at least two instruments must be relevant. But power of these tests will be low if K-1 instruments are weak.

instrument is 29.08 in the recent sample and 62.33 in the long sample. Thus, the instrument is strong by conventional standards in both cases.

Table 6—1st Stage and Reduced Form Results - PSID

	1997-2019 Sample		1975-2019 Sample	
	2SLS 1^{st} Stage	Reduced Form	2SLS 1^{st} Stage	Reduced Form
Dependent Variable:	ΔW	ΔH	ΔW	ΔH
Education	.0028	.0018	.0028	.0023
	(.0006)	(.0007)	(.0004)	(.0005)
	[.0005]	[.0005]	[.0004]	[.0004]
F-Stat (Hetero-σ Robust)	22.02	5.66	46.2	21.21
p-value	.000	.017	.000	.000
F-Stat (Cluster Robust)	29.08	10.33	62.33	40.00
p-value	.000	.001	.000	.000
Partial R^2	.0006	.0002	.0007	.0003

Note: Heteroskedasticity robust standard errors are in parentheses. Cluster robust standard errors (by individual) are in square brackets. All regressions control for year effects, age, age^2*, marriage status, and race/ethnicity.*

The second stage 2SLS results are reported in Table 7. In the recent sample the OLS estimate is -.287 while the 2SLS estmate is .621, so this pattern is very similar to the NLSY97 results. But here the heteroskedasticity and cluster robust t-stat results disagree; they are 1.93 (p=.053) and 2.52 (p=.012). But turning to the reduced form for hours results in Table 6, we see that the heteroskedasticity and cluster robust AR test p-values are .017 and .001 respectively. So the evidence for a significant positive Frisch elasticity is clear according to the AR test. When we invert the AR test we obtain a 95% confidence interval of (0.21, 1.28). Comparing this to the t-test confidence interval of (0.14, 1.10), we see it is further from zero but also asymmetric, with further extension in the positive direction.

In Online Appendix G we report a Monte Carlo experiment similar to that in Section V that shows the superior power properties of the AR test over the t-test are still substantial at this level of instrument strength. The AR test rejects the false null hypothesis $\beta = 0$ at a 67% rate, compared to 44% for the t-test.

In the long sample the 2SLS estimate of the Frisch elasticity increases to 0.88. Due to the large first-stage F, driven by the large sample size, the estimate is now fairly precise. The heteroskedasticity and cluster robust t-stats are 3.46 (p=.001) and 4.60 (p=.000). As we see in Table 6, AR test p-values are both .000. The AR 95% confidence interval is (0.52, 1.25), compared to the t-test confidence interval of (0.47, 1.17). These results illustrate how AR and t-test inference start to converge once the first-stage F-statistic is as large as 62.

Table 7—Frisch Elasticity Estimates - PSID

	1997-2019 Sample		1975-2019 Sample	
	OLS	2SLS	OLS	2SLS
Wage Change	-.287 (.008) [.008]	.621 (.321) [.246]	-.266 (.006) [.006]	0.882 (.237) [.178]
N	34,670		64,701	

Note: Heteroskedasticity robust standard errors are in parentheses. Cluster robust standard errors (by individual) are in square brackets. All regressions control for year effects, age, age^2*, marriage status, and race/ethnicity.*

Finally, Table 8 presents estimates from over-identified models that use education and its interactions with age and age^2 as excluded instruments. We report only GMM-CU and CLR test results as we argue that is the preferred approach. In the recent sample the first stage F is 12.5, and the Frisch elasticity estimate is .819, with a CLR p-value of .000 and confidence interval of (.320,1.586). The instruments also pass the J test. In the long sample the Frisch elasticity estimate increases to 1.019, but instruments fail to pass the J-test, so we ignore this result.

Table 8—Overidentified Frisch Elasticity Estimates - PSID

	1997-2019 Sample	1975-2019 Sample
Wage Change	.819 (.319) [.249]	1.019 (.209) [.160]
CLR-GMM Test	14.07	60.8
p-value	.000	.000
Olea-Pflueger Effective F	12.51	35.41
p-value	.000	.000
J Test	3.403	9.751
p-value	.1824	.008
N	34,670	64,701

Note: Heteroskedasticity robust standard errors are in parentheses. Cluster robust standard errors (by individual) are in square brackets. All regressions control for year effects, age, age^2*, marriage status, and race/ethnicity. The three excluded instruments are years of education: alone, interacted with age, and interacted with age*^{2}*.*

IX. Conclusion

The magnitude of the Frisch labor supply elasticity – how work hours respond to predictable wage changes – lies at the center of many economic policy debates, because the pure substitution effect measured by the Frisch elasticity is a vital input into tax policy. For example, a higher value of the Frisch elasticity implies a lower optimal tax rate on labor income. Because of its importance, there is a large literature estimating the Frisch elasticity using instrumental variable methods.

Classic studies that attempted to estimate the Frisch elasticity typically found it to be small and insignificant. But these studies suffered from two key problems: First, they were plagued by weak instrument problems, as it is hard to find instruments that are good predictors of wage growth. Second, they relied on estimation methods and inferential procedures (2SLS and the associated t-test) that are biased towards finding the Frisch elasticity is small and insignificant.

We revisit this issue, focusing on two improvements: First, our main instrument for wage growth is the ASVAB ability test, which is a better predictor of wage growth than the education and age variables used in classic studies. It generates a first-stage F statistic of 10.12, which exceeds conventional thresholds for an acceptably strong instrument. Second, we rely on inferential procedures that are robust to weak instrument problems and have better power properties.

We estimate a fairly large Frisch elasticity of 0.597 for young men using data from the NLSY97. But, as is typical of this literature, the 2SLS standard error is 0.403, implying our estimate is very imprecise. As a result, a 2SLS t-test cannot reject the hypothesis that the elasticity is zero at conventional levels – a result that is typical of many prior papers. However, we present Monte Carlo and analytical results showing the t-test is a very poor guide to inference in this context.

We show that the 2SLS t-test has little power to detect a true positive Frisch elasticity due to a strong *positive* association between 2SLS estimates and their standard errors that arises when the OLS bias is *negative* – as it is here. This causes positive estimates of the Frisch elasticity to have artificially inflated standard errors. In fact, we show that the power of the t-test is so poor that it has only about a 5.1% chance of detecting a true Frisch elasticity as large as 0.597. The power asymmetry that afflicts the t-test (poor power to detect effects opposite in sign to the OLS bias) was first noted in Keane and Neal (2023).[34]

Fortunately, the AR test of Anderson and Rubin (1949) largely avoid the power asymmetry problem. A Monte Carlo experiment shows that in our data environment the AR test has ten times the power of the t-test to detect a positive Frisch elasticity. The AR test indicates that our estimate of the Frisch elasticity is significant at the 3.5% or 1.8% level (depending on the level clustering).

[34] We suspect t-test power asymmetry has been widely ignored because it is unusual to study power of tests *conditional* on estimates, and conventional power curves do not reveal the sign pattern of rejections. In fact, Angrist and Kolesár (2021) reject the whole concept, stating "[the asymmetry] does not make conventional frequentist inference unreliable. The conventional standard for reliability of inference is the accuracy of confidence interval coverage, gauged without conditioning on parameter estimates."

Given the theoretical restriction that the Frisch elasticity is positive, we also consider one-tailed tests. This yields similar results: e.g., using heteroskedasticity robust standard errors, the 2SLS t-stat (1.48) indicates significance at the 6.9% level, while the AR test (t_{AR}=2.11) implies significance at the 1.7% level. Relative to a N(0,1), the density of the 2SLS t-test is highly skewed left and has very little mass in the right tail, so a one-tailed t-test against H_0:$\beta \leq 0$ has very poor power. Given its highly normal distribution, it would be more appropriate to call it an "ersatz t-test." In contrast, we show that, given a true $\beta = 0$, the distribution of t_{AR} is almost exactly a $N(0, 1)$ random variable in our data environment.

Theorists commonly recommend using AR rather than the t-test when instruments are weak; see, e.g., Andrews, Stock and Sun (2019). This is because, unlike the t-test, the AR test is robust to weak instrument problems, meaning it always has correct size. The new twist here is our claim the AR test has far superior power properties to the t-test even in contexts where instrument strength is well above conventional weak IV test thresholds (like the F>10 rule of thumb).

Our estimated Frisch elasticity of 0.597 for young men is large compared to the median estimate of 0.17 across 11 classic studies of men surveyed in Keane (2011). It is more in line with later studies by Lee (2001) and Ziliak and Kniesner (2005) who obtain estimates of 0.50 to 0.54. More recent studies surveyed in Keane (2021) present accumulating evidence that the Frisch elasticity increases substantially with age.[35] The seven studies surveyed give a mean (median) estimate of 0.58 (0.45) for young men, which is similar to our estimate here.[36]

We also estimate over-identified models that use the ASVAB, a numerical speed test, and the NLSY97 delinquency score as instruments for wage growth. In the over-identified case, two-step GMM suffers from the same association between estimates and standard errors as 2SLS, rendering t-tests unreliable. To avoid this problem, we recommend using LIML or continuously updated GMM, in conjunction with the conditional likelihood ratio (CLR) test. CLR is robust to weak instruments, avoids the power asymmetry of the t-test, and is more efficient than AR in the over-identified case.

Our GMM-CU estimate of the Frisch elasticity is .552, which is very similar to what we obtain using the single ASVAB instrument. But using multiple instruments increases efficiency, as the CLR p-value is .007. If we invert the CLR test we obtain a 95% confidence interval for the Frisch elasticity of 0.133 to 1.583, which covers most of the range that is typically used to calibrate macro models.

Finally, we show that our results generalize to the PSID, a much longer panel that contains men of all ages. Using PSID data on men aged 20-65 from 1997-2019, we obtain a Frisch estimate of .621. The t-test evidence for significance of this estimate is mixed. But the AR test p-value is .017 and the AR 95% confidence interval is (0.21, 1.28).

[35]See Imai and Keane 2004, Borella, De Nardi and Yang 2019, Erosa, Fuster and Kambourov 2016, French 2005, French and Jones 2012, Iskhakov and Keane 2021, and Keane and Wasi 2016.

[36]This increases substantially to a mean (median) of 1.56 (1.45) for 60 year-old men.

There have been several attempts to reconcile the small and insignificant 2SLS estimates of the Frisch elasticity obtained in classic micro data studies with the larger values used in macro calibrations. As Keane and Rogerson (2012, 2015) discuss, these fall into two broad categories: One set of explanations, exemplified by Imai and Keane (2004) and Domeij and Floden (2006) argues that estimation of equation (1) gives downward biased estimates of the Frisch elasticity due to problems created by human capital accumulation or liquidity constraints. The other set of explanations, exemplified by Chang and Kim (2006) and Rogerson and Wallenius (2009), argue that, once one accounts for the participation margin of labor supply and aggregation issues, it is possible for the macro level Frisch elasticity to be large even if the micro level elasticity is small.[37]

Our argument here is complementary but new in that we criticise the micro-econometric literature on its own terms: Suppose the assumptions necessary for a 2SLS regression of hours changes on wage changes to deliver consistent estimates of the Frisch elasticity do in fact hold. Even then, we show that the econometric methods that have been used to draw inferences from those estimates are inherently biased against finding the elasticity is both large and significant.

The power asymmetry problem that afflicts the 2SLS t-test is relevant beyond the Frisch elasticity application. In a different context, where the OLS bias is positive, 2SLS standard errors on *positive* estimates would be spuriously precise, making it difficult for 2SLS t-tests to detect a true *negative*. In the classic application of instrumental variables to estimate a treatment effect given positive selection into treatment, 2SLS t-tests have difficulty detecting true negative effects, violating a "first do no harm" principle in policy evaluation.

Furthermore, the association between 2SLS estimates and their standard errors that generates the t-test power asymmetry problem vanishes extremely slowly as instrument strength increases – see Keane and Neal (2022). Thus, we argue that researchers ought to adopt the AR test or CLR test in lieu of the t-test even when instruments are quite strong.

[37] More recently, Gottlieb, Onken and Valladares-Esteban (2021) show a large Frisch elasticity can be reconciled with modest reactions to tax holidays due to a combination of income and equilibrium effects.

ACKNOWLEDGEMENTS

We thank Peter Phillips, Josh Angrist, Michal Kolesar and Isaiah Andrews for helpful comments. This research was supported by Australian Research Council (ARC) grants DP210103319 and CE170100005.

REFERENCES

Altonji, J.G. 1986. "Intertemporal substitution in labor supply: Evidence from micro data." *Journal of Political Economy*, 94(3, Part 2): S176–S215.

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

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

Angrist, Joshua, and Michal Kolesár. 2021. "One instrument to rule them all: The bias and coverage of just-id iv." National Bureau of Economic Research.

Borella, M., M. De Nardi, and F. Yang. 2019. "Are marriage-related taxes and Social Security benefits holding back female labor supply?" National Bureau of Economic Research.

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

Browning, Martin, Angus Deaton, and Margaret Irish. 1985. "A profitable approach to labor supply and commodity demands over the life-cycle." *Econometrica: journal of the econometric society*, 503–543.

Chang, Y., and S. Kim. 2006. "From individual to aggregate labor supply: A quantitative analysis based on a heterogeneous agent macroeconomy." *International Economic Review*, 47(1): 1–27.

Conesa, J.C., S. Kitao, and D. Krueger. 2009. "Taxing capital? Not a bad idea after all!" *American Economic Review*, 99(1): 25–48.

Domeij, D, and M. Floden. 2006. "The labor-supply elasticity and borrowing constraints: Why estimates are biased." *Review of Economic dynamics*, 9(2): 242–262.

Erosa, A., L. Fuster, and G. Kambourov. 2016. "Towards a micro-founded theory of aggregate labour supply." *The Review of Economic Studies*, 83(3): 1001–1039.

French, E. 2005. "The effects of health, wealth, and wages on labour supply and retirement behaviour." *The Review of Economic Studies*, 72(2): 395–427.

French, Eric, and John Jones. 2012. "Public pensions and labor supply over the life cycle." *International Tax and Public Finance*, 19(2): 268–287.

Gottlieb, C., J. Onken, and A. Valladares-Esteban. 2021. "On the Measurement of the Elasticity of Labour." *European Economic Review*, 139.

Hansen, Lars Peter, John Heaton, and Amir Yaron. 1996. "Finite-sample properties of some alternative GMM estimators." *Journal of Business & Economic Statistics*, 14(3): 262–280.

Imai, S., and M.P. Keane. 2004. "Intertemporal labor supply and human capital accumulation." *International Economic Review*, 45(2): 601–641.

Iskhakov, Fedor, and Michael Keane. 2021. "Effects of taxes and safety net pensions on life-cycle labor supply, savings and human capital: The case of Australia." *Journal of Econometrics*, 223(2): 401–432.

Keane, Michael, and Timothy Neal. 2023. "Instrument strength in IV estimation and inference: A guide to theory and practice." *Journal of Econometrics*, 235(2): 1625–1653.

Keane, Michael P, and Nada Wasi. 2016. "Labour supply: the roles of human capital and the extensive margin." *The Economic Journal*, 126(592): 578–617.

Keane, M.P. 2011. "Labor supply and taxes: A survey." *Journal of Economic Literature*, 49(4): 961–1075.

Keane, M.P. 2021. "Recent Research on Labor Supply: Implications for Tax and Transfer Policy." *Labour Economics*, 102026.

Keane, M.P., and R. Rogerson. 2012. "Micro and macro labor supply elasticities: A reassessment of conventional wisdom." *Journal of Economic Literature*, 50(2): 464–76.

Keane, M.P., and R. Rogerson. 2015. "Reconciling micro and macro labor supply elasticities: A structural perspective." *Annu. Rev. Econ.*, 7(1): 89–117.

Keane, M.P., and T. Neal. 2022. "A Practical Guide to Weak Instruments." UNSW Economics Working Paper No. 2021-05a. Available at SSRN: https://ssrn.com/abstract=3846841.

Kleibergen, Frank. 2005. "Testing parameters in GMM without assuming that they are identified." *Econometrica*, 73(4): 1103–1123.

Lee, Chul-In. 2001. "Finite sample bias in IV estimation of intertemporal labor supply models: is the intertemporal substitution elasticity really small?" *Review of Economics and Statistics*, 83(4): 638–646.

Lee, David S, Justin McCrary, Marcelo J Moreira, and Jack Porter. 2022. "Valid t-ratio Inference for IV." *American Economic Review*, 112(10): 3260–90.

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

Moreira, Humberto, and Marcelo J Moreira. 2019. "Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors." *Journal of Econometrics*, 213(2): 398–433.

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

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

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

Phillips, Peter CB. 1989. "Partially identified econometric models." *Econometric Theory*, 5(2): 181–240.

Prescott, E.C. 2006. "Nobel lecture: The transformation of macroeconomic policy and research." *Journal of Political Economy*, 114(2): 203–235.

Rogerson, R., and J. Wallenius. 2009. "Micro and macro elasticities in a life cycle model with taxes." *Journal of Economic theory*, 144(6): 2277–2292.

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

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

Stock, J., and M. Wright. 2000. "GMM with Weak Identification." *Econometrica*, 68(5): 1055–96.

Stock, J., and M. Yogo. 2005. "Testing for weak instruments in linear IV regression." *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 80–108.

Ziliak, James P, and Thomas J Kniesner. 1999. "Estimating life cycle labor supply tax effects." *Journal of Political Economy*, 107(2): 326–359.

Ziliak, James P, and Thomas J Kniesner. 2005. "The effect of income taxation on consumption and labor supply." *Journal of Labor Economics*, 23(4): 769–796.

Appendix: Analytical Power Calculations for the AR and t-tests

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

$$\text{(A1)} \qquad \begin{gathered} y = \beta x + u \\ x = \pi z + e \quad \text{where} \quad e = \rho u + \sqrt{1-\rho^2}\eta \\ u \sim iidN(0,1), \quad \eta \sim iidN(0,1), \quad z \sim iidN(0,1) \end{gathered}$$

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

$$\text{(A2)} \qquad Power_{AR}(\beta|\lambda,\rho) = \Phi(\lambda D - z_{1-\alpha/2}) + \Phi(-z_{1-\alpha/2} - \lambda D)$$

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

A1. The Infeasible IV and AR tests have Identical Power, so AR is Optimal

A simple way to understand why the AR test is optimal (i.e., the most powerful unbiased test) is to note that its power curve is identical to that of the infeasible IV (IIV) estimator one could construct if, hypothetically, one knew the true first-stage π and did not have to estimate it. From (A1) we have:

$$\text{(A3)} \qquad y = \beta(\pi z) + v \quad \text{where} \quad v = \beta e + u$$

It is simple to derive the power curve for the IIV estimator, as IIV is simply the OLS estimator of β one obtains by regressing y on πz. First, note that:

$$\text{(A4)} \qquad se(\hat{\beta}_{IIV}) = \sigma_v/\sqrt{n}\pi\sigma_z \quad \text{where} \quad \sigma_v = \sqrt{\beta^2\sigma_e^2 + 2\beta cov(e,u) + \sigma_u^2}$$

Then define $W = \hat{\beta}_{IIV}/se(\hat{\beta}_{IIV})$ and $M = \beta/se(\hat{\beta}_{IIV})$ so that $W \sim N(M,1)$ in large samples. The power of the t-test for significance of $\hat{\beta}_{IIV}$ is simply the probability that $|W|$ exceeds the relevant critical value, expressed as a function of true β. For example, for a 5% level two-tailed test, we have:

$$\text{(A5)} \qquad Power_{IIV}(\beta) = \Phi(M - 1.96) + \Phi(-M - 1.96)$$

Notice that if we impose all the variance normalizations in (A1) we then have:

$$\text{(A6)} \qquad M = \beta\sqrt{N}\pi\sigma_z/\sigma_v = \sqrt{N}\pi\left(\beta/\sqrt{\beta^2+2\beta\rho+1}\right) = \lambda D$$

Thus (A6) is equivalent to (A2). So the AR and IIV power curves are identical.

This equivalence may seem remarkable, as IIV assumes π is known. But AR gives equivalent power because it provides a way to test for significance of $\hat{\beta}_{2SLS}$ that doesn't require knowledge of π. The IIV approach relies on a regression of y on πz, while the AR approach relies on the reduced form regression of y on z or, equivalently, the regression of y on $\hat{\pi} z$. All three regressions yield the exact same t-test, as multiplication of z by a constant like π or $\hat{\pi}$ does not alter its predictive ability for y. It only alters the scale the slope coefficient on z.

The advantage of knowing π is that it would, hypothetically, allow one to obtain a more precise estimate of β, as one knows the correct scaling. But it does not alter inference about whether x has an effect on y, as the IV assumptions mean x has a significant effect on y if and only if z has a significant on y.

A2. Power of the t-Test

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

$$\text{(A7)} \quad Power_t(\beta|\lambda,\rho) = \int_{-\infty}^{\infty} \left(\mathbb{I}\{t^2 \geq (1-\rho_0^2) z_{1-\alpha/2}^2\} f(t,D,\lambda,\rho_0) + \mathbb{I}\{t^2 \geq z_{1-\alpha/2}^2\} \right) \phi(t-\lambda) dt$$

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

$$\text{(A8)} \quad f(t,D,\lambda,\rho_0) = \Phi\left(\frac{a_2 - \lambda D - \rho_0(t-\lambda)}{\sqrt{1-\rho_0^2}}\right) - \Phi\left(\frac{a_1 - \lambda D - \rho_0(t-\lambda)}{\sqrt{1-\rho_0^2}}\right),$$

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

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

The integral in (A7) must be evaluated numerically.

A3. Construction of Figure 3 in the Main Text

To form Figure 3 in the text, we set $\alpha = 0.05$, and set $\lambda = 3.186$ and $\rho = -0.7$ to mimic the empirical results in Section V, as $\lambda = 3.186$ corresponds to a true F of 10.12. The power asymmetry of the t-test is evident in Fig. 3, as it has little power to detect a wide range of true positive β values because the OLS bias is negative ($\rho = -0.70$). The power curve for IIV is identical to that of AR.

A4. Explaining the Power Asymmetry of the T-test

Here we give a simple explanation of the power asymmetry in the t-test. The standard error of $\hat{\beta}_{2SLS}$ is $\hat{\sigma}/\sqrt{N\hat{cov}(z,x)^2/\hat{\sigma}_z^2}$. First, $\hat{\sigma}^2$ is a quadratic minimized when $\hat{\beta}_{2SLS} = \hat{\beta}_{OLS}$. Second, $\hat{cov}(z,x)$ tends to be larger when $\hat{\beta}_{2SLS}$ is shifted towards OLS. To see this, recall that the 2SLS estimator of β is given by:

$$\text{(A9)} \qquad \hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i x_i} = \beta + \frac{\sum_{i=1}^{n} z_i u_i}{\sum_{i=1}^{n} z_i x_i} = \beta + \frac{\widehat{cov}\,(z,u)}{\widehat{cov}\,(z,x)}$$

We assume without loss of generality that the population covariance between the instrument and the endogenous variable is positive, $cov(z,x) > 0$. To simplify, we further assume that $\widehat{cov}(z,x) > 0$, so the sign of the coefficient on z in the first-stage regression is correct. Violation of this condition is extremely rare if the instrument is reasonably strong. For instance, in our Monte Carlo experiment in Section V, where F=10.12, first-stage sign is correct in all $10,000$ replications.

Given that $\widehat{cov}(z,x) > 0$, equation (A9) makes clear that the sign of $\widehat{cov}(z,u)$, the sample covariance between the instrument and the structural error, determines whether $\hat{\beta}_{2SLS}$ lies above or below the true β.[38] Recall that the sign of ρ determines the sign of the OLS bias. Thus, if $\rho\widehat{cov}(z,u)$ is positive the 2SLS estimate is shifted towards the OLS bias, and vice versa.

Our key point is that a larger sample realization of $\rho\widehat{cov}(z,u)$ also drives up the sample covariance between the instrument and the endogenous variable, making the instrument appear spuriously strong, as is obvious because:

$$\text{(A10)} \qquad \widehat{cov}\,(z,x) = \pi\widehat{var}\,(z) + \rho\widehat{cov}\,(z,u) + \sqrt{1-\rho^2}\,\widehat{cov}\,(z,\eta)$$

This spurious instrument strength drives down the 2SLS standard error.

Thus, a positive sample realization of $\rho\widehat{cov}(z,u)$ generates an estimate shifted towards OLS. It also generates a low standard error because a large $\rho\widehat{cov}(z,u)$ leads to a large $\widehat{cov}(z,x)$. Hence, 2SLS will appear spuriously precise in samples where the estimated coefficient is most shifted in the direction of the OLS bias. Conversely, estimates shifted away from OLS appear spuriously imprecise. This generates the power asymmetry in the 2SLS t-test.

A5. A Brief Explanation of Weak Instrument Tests

The widely used Stock and Yogo (2005) weak IV tests assess whether instruments are strong enough for t-test size distortions to be modest. These tests are derived using the t-test power function in equation (A7). The size of a test is defined as the probability of rejecting H_0:$\beta = 0$ when the null hypothesis is true. Thus, the size of the t-test is simply the power function evaluated at β=0.

[38]It follows that 2SLS is approximately median unbiased provided the instrument is strong enough that an incorrect first-stage sign a rare event, as only such events impart median bias.

A problem arises because the power function also depends on the degree of endogeneity ρ. To sidestep this issue, Stock and Yogo (2005) focus on the maximum (on the high side) size that arises when endogeneity is very severe, $\rho = \pm 1$. For example, by setting β=0 and $\rho = \pm 1$ in the power expression in equation (A7), and doing a grid search for the level of $F = \lambda^2$ that sets power approximately equal to 15%, they obtain F=1.82.[39] Thus, if the first-stage F is 1.82 a 5% level two-tailed t-test will reject a true null hypothesis at a 15% rate, giving a size distortion of 10%. This is the maximal or "worst-case" size distortion.

A second problem arises as in any given sample we cannot observe the true $F = N \cdot Var(z\pi)/\sigma_e^2 = N\pi^2\sigma_z^2/\sigma_e^2$. Rather, we can only observe the sample realization $\hat{F} = N\hat{\pi}^2\hat{\sigma}_z^2/\hat{\sigma}_e^2$. The sample $\hat{F}$ is a draw from the non-central F-distribution with non-centrality parameter F. For instance, in the single instrument case, a sample $\hat{F} > 8.96$ gives 95% confidence that F is at least 1.82.

Stock and Yogo (2005) weak IV tests give sample $\hat{F}$ thresholds that give 95% confidence that true first-stage F is above some threshold, where that F in turn implies a certain maximal size distortion. For example, a sample $\hat{F} > 8.96$ gives 95% confidence that F is at least 1.82, which in turn, implies the maximum size of a two-tailed 5% t-test is 15% (i.e., maximal size distortion of 10%).[40]

Both Keane and Neal (2023) and Angrist and Kolesár (2021) have criticized the focus on the worst-case scenario of ρ=±1, arguing that for most plausible levels of endogeneity the size distortion is much less. But in addition, we also criticize the exclusive focus on size to the neglect of broader power considerations.

A6. Power of the AR vs t-test at Different Levels of Instrument Strength

A critical point is that weak IV tests do not reveal if the estimator has acceptable power properties. We now present some comparisons of the power properties of the AR and t-tests at different levels of instrument strength F.

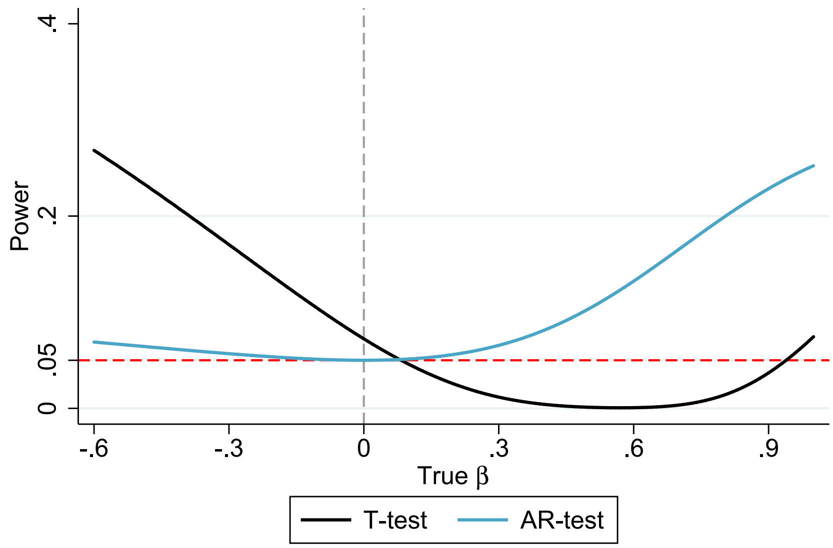
Figure A1 presents the case of F=1, which is indicative of the poor instrument strength in classic studies of the Frisch elasticity. We see the 2SLS t-test has essentially no power to detect positive Frisch elasticities in the plausible range of 0.1 to 0.9 (power is less than the 5% size of the test throughout this range). The AR test performs better: It does have power greater than size at all levels of β except $\beta = 0$, reflecting that it is an unbiased test. But its power is still very low: It doesn't pass 20% until the elasticity exceeds 0.8. This reflects the fact that the data is simply not very informative at this low level of instrument strength.

Another notable feature of Figure A1 is that the t-test appears to have much better power than the AR test for negative values of true β. This reveals the flip side of the power asymmetry problem: In samples where the 2SLS estimate

[39] Thus, the figure of 1.82 is subject to numerical error, and should not be viewed as exact. The same is true of other weak IV thresholds in this literature.

[40] Similarly, a sample $\hat{F}$ of 10 gives 95% confidence that true F is at least 2.3, which implies a maximal size distortion of 8.5%. And a sample $\hat{F} > 16.4$ gives 95% confidence that F is at least 5.78, which in turn implies the size distortion is no more than 5%.

FIGURE A1. POWER OF THE T-TEST VS. AR-TEST WHEN TRUE $F = 1$ ($\rho = -0.7$)



is shifted in the direction of OLS, which in this case means it is shifted in the negative direction, the 2SLS standard error is spuriously small, which inflates the power of the t-test. This is not a desirable property, as the standard error exaggerates the precision of the estimate in such cases.

Figure A2 considers the case of F=2.3. This is particularly interesting, as a sample $\hat{F}$ of 10 is required to give 95% confidence that F is at least 2.3. Thus, this case corresponds to the widely used Staiger-Stock rule of thumb, that a first-stage $\hat{F}$ of at least 10 indicates an acceptable level of instrument strength. However, Figure A2 reveals that the t-test has very poor properties in this case. It has essentially no power to detect positive Frisch elasticities in the plausible range of 0.1 to 0.8, as power is less than the 5% size of the test throughout this range. The AR test has much better power to detect a true positive Frisch elasticity, but its power is still uninspiring (e.g., it doesn't pass 20% until the elasticity exceeds 0.5). So the data is not very informative at this level of instrument strength.

The severe bias of the t-test is also evident in Figure A2. Power is minimized in the vicinity of β=0.50 rather than at β=0. This again reflects the power asymmetry pf the t-test, and the fact that it has little power to detect a wide range of plausible positive elasticities because the OLS bias is negative.

Figure 3 in the main text considers the case of F=10.12, which is the value we used in the Monte Carlo exercise in Section V. A first-stage $\hat{F}$ of at least 23.2 is required to have 95% that true F is at least 10.12. In this case the t-test has power less than size for true effects in the 0.05 to 0.50 range, so, as we note in the text, it is uninformative over that range. The power of the AR test is far superior, reaching about 60% when true elasticity is 0.50.

FIGURE A2. POWER OF THE T-TEST VS. AR-TEST WHEN TRUE $F = 2.3$ ($\rho = -0.7$)

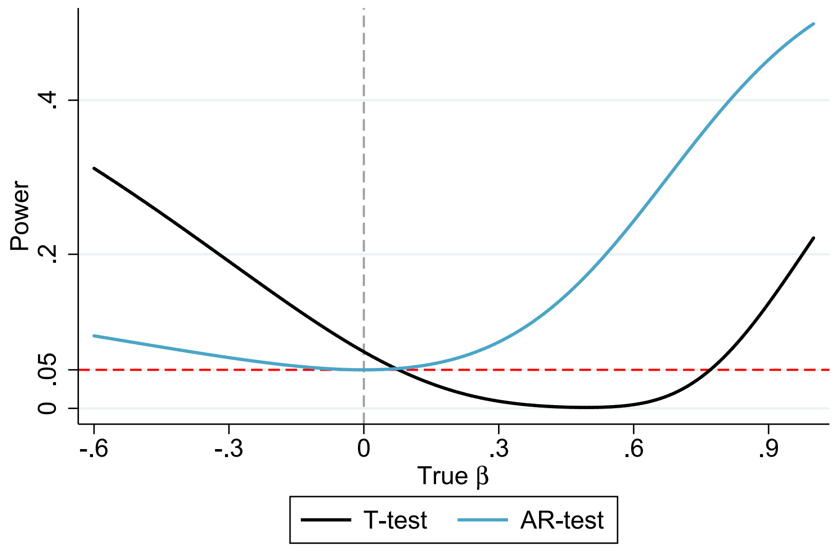
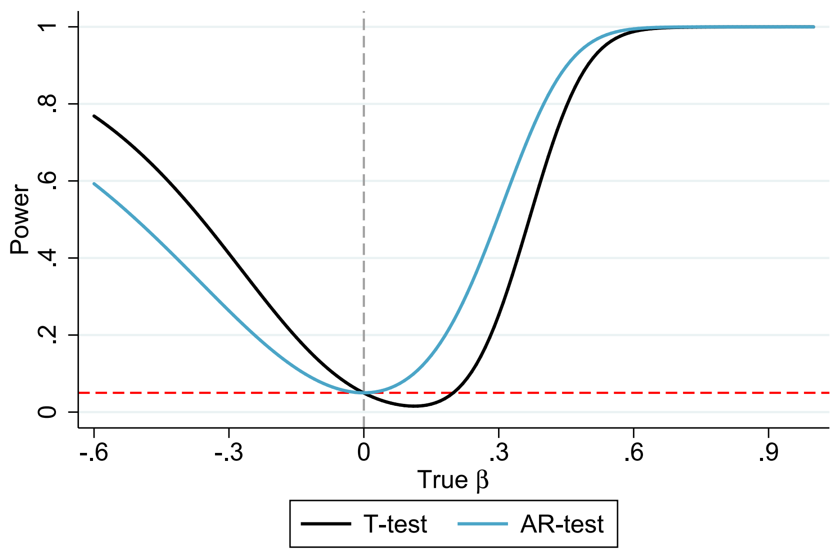
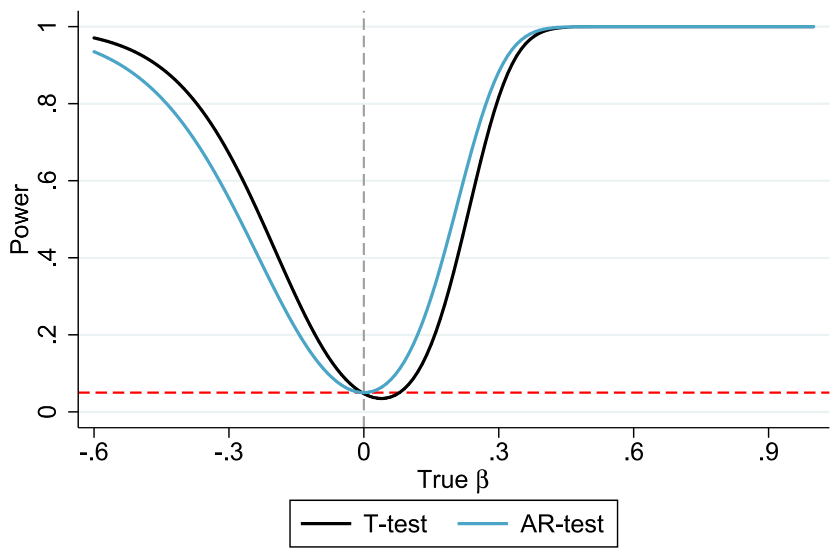


FIGURE A3. POWER OF THE T-TEST VS. AR-TEST WHEN TRUE $F = 29.44$ ($\rho = -0.7$)



An obvious question is how large F must be for the t-test to begin to exhibit acceptable power for plausible elasticity values. Figure A3 reports results for a true F of 29.44. A first-stage $\hat{F}$ of at least 50 is required to have 95% confidence that true F is at least this large. At this level of instrument strength the power of both tests approaches one when the true elasticity approaches 0.6. However, the t-test still has power less than size for elasticities in the 0.0 to 0.2 range, and very poor power compared to the t-test for elasticities in the 0.0 to 0.4 range. The size of the t-test (i.e., power at $\beta = 0$) is 4.96% in this case, so it is very close to the correct 5%. But bias is still evident as power is minimized at an elasticity of roughly $\beta = 0.1$.

FIGURE A4. POWER OF THE T-TEST VS. AR-TEST WHEN TRUE $F = 73.75$ ($\rho = -0.7$)



Finally, Figure A4 considers the case of true F=73.75, which is a very high level of instrument strength. A first-stage $\hat{F}$ of at least 104.7 is required to have 95% confidence that true F is at least this large. We examine this case because Lee et al. (2022) show that a first-stage sample $\hat{F}$ of at least 104.7 is required for the worst-case size distortion in the t-test to be no more than 5%.[41]

At this high level of instrument strength the power curves of the two tests are much more similar, and power of both tests approaches 1 for β around 0.40. But the power advantage of the AR test is still evident in the $\beta \in (0.0, 0.30)$ range. For instance, for β=0.15 the AR test power is 40% vs. 25% for the t-test.

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

[41]Their analysis is subtly different from Stock and Yogo (2005), in that their "worst case" refers to the maximum size distortion over all possible values of endogeneity ρ and all possible values of the true F. The worst case scenario for ρ is again ± 1, while the worst case for F is $[\hat{F}/(\sqrt{\hat{F}} + 1.96)]^2$.