ARC CENTRE OF EXCELLENCE IN
POPULATION AGEING RESEARCH

ARC Centre of Excellence in Population Ageing Research

Working Paper 2021/25

Robust Inference for the Frisch Labor Supply Elasticity

Michael Keane and Timothy Neal

Robust Inference for the Frisch Labor Supply Elasticity

By Michael Keane and Timothy Neal*

October 6, 2021

There is a long standing controversy over the magnitude of the Frisch labor supply elasticity. Macro economists using DSGE models often calibrate it to be large, while many micro data studies find it is small. Several papers attempt to reconcile the micro and macro results. We offer a new and simple explanation: Most micro studies estimate the Frisch using a 2SLS regression of hours changes on income changes. But available instruments are typically "weak." In that case, we show it is an inherent property of 2SLS that estimates of the Frisch will (spuriously) appear more precise when they are more shifted in the direction of the OLS bias, which is negative. As a result, Frisch elasticities near zero will (spuriously) appear to be precisely estimated, while large estimates will appear to be imprecise. This pattern makes it difficult for a 2SLS t-test to detect a true positive Frisch elasticity. We show how the use of a weak instrument robust hypothesis test, the Anderson-Rubin (AR) test, leads us to conclude the Frisch elasticity is large and significant in the NLSY97 data. In contrast, a conventional 2SLS t-test would lead us to conclude it is not significantly different from zero. Our application illustrates a fundamental problem with 2SLS t-tests that arises quite generally, even with strong instruments. Thus, we argue the AR test should be widely adopted in lieu of the t-test.

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

JEL: *J22, D15, C12, C26*

I. Introduction

The elasticity of labor supply with respect to predictable wage changes – known as the Frisch elasticity – plays a key role in many economic policy debates. The Frisch is special because predictable wage changes have pure substitution effects. As an example of its importance, Conesa, Kitao and Krueger (2009) argue that higher values of the Frisch imply a higher optimal tax rate on capital income. And macro models where real shocks play a key role in business cycles often require the Frisch 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 exemplified by classic papers by MaCurdy (1981) and Altonji (1986) and surveyed in Keane (2011, 2021).

* CEPAR & School of Economics, UNSW. Corresponding author: m.keane@unsw.edu.au

Classic micro data studies in the style of MaCurdy (1981) typically find the Frisch elasticity is small, while macro economists using DSGE models often calibrate it to be large. This led to a long-standing "macro-micro controversy" over the magnitude of the Frisch. Keane and Rogerson (2012, 2015) discuss attempts to resolve the controversy. Here we present a new type of resolution based on a critique of the micro-econometrics itself: We argue the classic studies were inherently biased against finding the Frisch is both large and significant, due to weak instrument problems and a little appreciated generic property of 2SLS t-tests.

The basic idea behind most micro studies is as follows: Given panel data on workers observed over time, one may run an OLS regression of changes in log work hours on changes in log wages. But this fails to deliver the elasticity of hours with respect to *predictable* wage changes – as some wage changes are surprises. Instead, the approach pioneered by MaCurdy (1981) involves running a 2SLS regression where one instruments for the change in log wages using an instrument with two 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 surprise wage changes). Then, fitted values from the first-stage 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.

Unfortunately, the literature on estimating the Frisch elasticity has been hampered by weak instrument problems. This is because it is hard to find variables that are both known in advance and are good predictors of a person's wage growth during the next year. In other words, the lion's share of annual wage growth at the individual level appears to be idiosyncratic or unpredictable.

A little appreciated property of 2SLS is that it generates a strong association between the 2SLS estimate and the standard error of regression, which is minimized when the 2SLS estimate is close to $E(\hat{\beta}_{OLS})$, see Phillips (1989). In Keane and Neal (2021) we show how this generates a strong association between the 2SLS estimates and their standard errors, and this association is positive if the OLS bias is negative. This makes it difficult for a 2SLS t-test to detect a true positive Frisch elasticity, especially if instruments are weak. Using NLSY97 data, we show how the Anderson-Rubin (AR) test, which is robust to weak instruments and less subject to this problem, leads us to conclude the Frisch elasticity is large and highly significant, while a conventional 2SLS t-test indicates it is not significantly different from zero. Thus, application of an appropriate inferential procedure – the AR test – reveals clear evidence to support a large Frisch elasticity.

It is important to emphasize that our critique of the classic micro studies is deeper than a claim that weak instruments make the micro estimates of the Frisch imprecise. Instead, we show it is an inherent property of 2SLS that estimates of the Frisch will (spuriously) appear more precise when they are more shifted in the direction of the OLS bias, which is negative. This will systematically bias micro data studies that rely on 2SLS t-tests against concluding the Frisch is large.

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.[1] We use data from rounds 11 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.[2]

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 instrument for wage growth. The motivation is that annual wage growth tends to be faster for more educated workers.[3] 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.[4] We find that the ASVAB percentile score is a stronger predictor of wage growth than education, so we use that as our instrument. But the idea is similar: Not surprisingly, wage growth is predictably faster for higher ability workers.

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

[2] 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%.

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

[4] The ASVAB measures aptitude in several areas including mathematics, general science, paragraph comprehension, and mechanical skills. It was administered in summer 1997 to spring 1998, when the youth were aged 13 to 18 (those aged 13 to 14 were given an easier version of the test). The NLS grouped respondent's into three-month age windows and calculated a youth's percentile rank within his age group.

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.[5] For this reason, 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 shows the results from estimating regressions of changes in log hours on changes in log wages, as in equation (1). The first column shows OLS results. The coefficient on the log wage change is -0.42 and very highly significant, with a standard error of 0.015.[6] This implies that a 10% wage increase is associated with a 4.2% reduction in hours of work. There are two reasons for a negative relationship: As we already noted, 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.

Another important factor driving the OLS estimate negative is the phenomenon of "denominator bias" that plagues many labor supply studies. The problem is that the wage rate is measured as the ratio of earnings to hours. If the hours variable in the denominator is measured with error, it causes a worker's measured wage to be too low precisely when his/her 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 look at the 2SLS results. The second column of Table 1 shows the first stage of 2SLS, where we regress log wages changes on the ASVAB percentile score to construct predictable wage changes. The coefficient is 0.039 and highly significant (standard error 0.012). This means a male worker in the 100th percentile of ability is predicted to have annual wage growth 3.9% 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. This implies significance at much better than the 1% level.

It is important to note, however, that the R^2 of the first stage regression is only .007, implying a correlation between our predictions and actual wage changes of .084. In fact, the partial R^2 that shows the fraction of wage variation explained by the ASVAB test alone is .002, implying a partial correlation of only .041.

[5] It is not clear if this is because wages grow relatively faster for high ability men than for high ability women, or because the ASVAB is not as good a proxy for labor market skills of women.

[6] 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	-0.416 (0.015) [0.015]		0.597 (0.403) [0.363]	
ASVAB Ability Score		0.039 (0.012) [0.011]		0.024 (0.011) [0.010]
F-Stat (Hetero-σ Robust)		10.12		4.47
p-value		0.002		0.035
F-Stat (Cluster Robust)		12.23		5.64
p-value		0.001		0.018
R^2	0.210	0.007		0.009

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

This illustrates the point that annual wage growth is very hard to predict. It is important to emphasize, however, that a higher R^2 in the first stage would not necessarily be a good thing. Measured wage changes contain both unpredictable and measurement error components that we specifically want to filter out, so we actually want the R^2 of the first stage regression to be much 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. Strikingly the estimate 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.

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

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.[7] This imprecision leaves us in a quandry over what we ought to conclude from the analysis.

[7] 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).

The imprecision in our 2SLS estimate is a consequence of the fact that the ASVAB score only explains a small part of the variance of wage changes. Because the partial correlation between the ASVAB score and wage changes is .041, the standard error goes up by a factor of 25 when we go from OLS to 2SLS (i.e., 1/.041 $\approx$ 25). This imprecision in 2SLS estimates has plagued the entire literature on estimating the Frisch elasticity using the 2SLS approach.

IV. An Example of the Weak Instrument Problem

The situation we see here, which is very typical of attempts to estimate the Frisch elasticity, is a classic example of the "weak instrument" problem. This refers to a situation where the instrument is statistically significant in the first stage of 2SLS, but it only explains a small part of the variance in the endogenous variable. In the present case, the ASVAB score is highly significant in the first stage (p=.002), but it it only explains a small part of the variance in wage changes (partial correlation = 0.04). It is statistically significant because even small effects tend to be significant when sample size is this large (N = 5,931).

Unfortunately, 2SLS results can be very unreliable when instruments are weak. In particular, 2SLS t-tests may be unreliable, and 2SLS estimates may be biased towards OLS. The important paper by Bound, Jaeger and Baker (1995) made applied economists acutely aware of these problems. This in turn led to an explosion of theoretical work on the "weak instrument problem." This work seeks to find criteria that instruments should satisfy for 2SLS results to be reliable.

The key insight of the weak instrument literature is that the quality of 2SLS estimates depends crucially on the size of the first stage partial F-statistic that tests significance of the instrument, where bigger is better. It is useful to recall the basic relationship that $F = NR^2/(1-R^2)$. Properties of 2SLS do not depend on N or first-stage R^2 *per se*, but only how they combine to form F. So a large sample size alone is not sufficient to ensure that 2SLS will deliver reliable results.

In an important paper, Staiger and Stock (1997) studied behavior of the 2SLS estimator at different levels of instrument strength. They developed the well-known "Staiger-Stock" rule of thumb, which says that the first-stage F should be at least 10 before we have confidence in 2SLS results. This $F > 10$ advice has been widely adopted in practice and presented in textbooks.[8,9] Of course, this is only meant as a rough guide, so a first stage-F near 10 puts one in a borderline case where weak instruments may or may not be a problem.

In our application to estimating the Frisch elasticity, the heteroskedasticity robust first-stage partial F-statistic for testing significance of the ASVAB in-

[8] 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."

[9] Stock and Yogo (2005) proposed critical values for F based on maximal size distortion in t-tests one is willing to tolerate. $F > 16.4$ ensures that a two-tailed 5% t-test will reject at a 10% rate or less. In other words, it has a size distortion of no more than 5%. But passing such a test does not imply the t-test will have acceptable power, as we illustrate in Keane and Neal (2021).

strument is 10.12. So we are right on the borderline between a “weak” and an acceptably strong instrument.[10] Should we trust the 2SLS results in this case? Is the 2SLS t-test, which tells us that the Frisch elasticity is not significantly different from zero, really reliable?

V. The Anderson-Rubin Approach

Early in the history of IV methods, Anderson and Rubin (1949) developed an alternative method that we can also use 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 means 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.

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 means 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 results are completely invalid anyway, and the t-test result is also meaningless. The very assumptions that make the IV approach valid in the first place 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-stat of 2.18 (p-value 0.035). So we are left with a quandry: The AR test says our 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 two 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 is unreliable: If instruments are weak, a 5% t-test may reject a true hypothesis at rates far above/below 5%,

[10] Andrews, Stock and Sun (2019) point out that *in general* it is inappropriate to use either a heteroskedasticity-robust or conventional F-test to assess instrument strength in non-homoskedastic settings, and suggest using the Olea and Pflueger (2013) effective first-stage F-statistic. However, as they point out, in the single instrument just-identified case that we consider here, this reduces to the conventional heteroskedasticity-robust F.

depending on details of the situation. Second, Moreira (2009) shows that in the case of a single instrument (which is what we have here) the AR test is the most powerful robust 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 robust test.[11]

Despite its clear advantages, the AR test has been widely neglected by applied researchers. In fact, as far as we know, it has never been adopted in the large literature on estimating the Frisch elasticity. In our Frisch elasticity application, given that the first-stage F statistic is only slightly above 10, conventional wisdom says we are in a borderline case where weak instruments may or may not be a concern. Clearly the AR test should be viewed as more reliable than the t-test in this context. It turns out the difference in performance between the two tests is not at all subtle. In the next section we present a numerical experiment based on our data that shows the AR test is *dramatically* superior in practice.

VI. Monte Carlo Experiment

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.[12]

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	3.8636	251.4631	6.4577	0.0111	0.0003

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

[11] Andrews, Stock and Sun (2019) argue that the AR test should be widely adopted by applied researchers. They state its advantages 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 in the usual sense and so does not sacrifice power relative to the conventional t-test. ... Since AR confidence sets are robust to weak identification and are efficient in the just-identified case, there is a strong case for using these procedures in just-identified settings."

[12] 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. OLS Estimates and Standard Errors

The first thing to notice in Table 2 is that both the median and mean of the OLS estimates of the Frisch elasticity (across all 10,000 datasets) are equal (to three decimal places) to the (downward biased) value of -0.416 we obtained using the original NLS sample. This is as expected, as our 10,000 artificial "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. In other words, the OLS standard error estimate is useful for making judgements about statistical significance.[13]

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 we obtained using the original NLS dataset. Again, this is exactly as expected. As all 10,000 of our artificial datasets were 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 (while of course OLS does not).

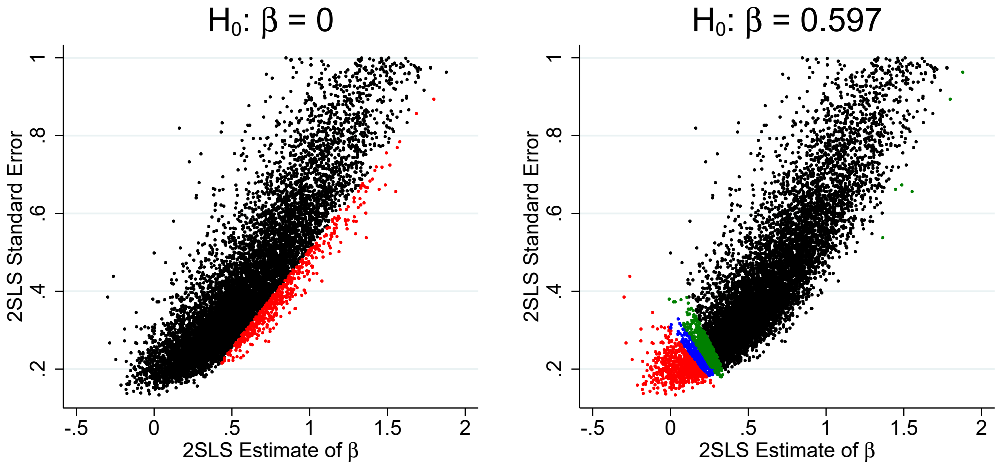
Second, note that the median of the estimated 2SLS standard errors, reported in the first row of Table 2, is 0.401. This agrees closely with the 2SLS standard error estimate of 0.403 in Table 1. However, the actual empirical standard deviation of the 2SLS estimates across the 10,000 data sets is 3.864. In contrast to OLS, this 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.[14] 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 2SLS estimates and their standard errors: The Spearman

[13] 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 we have an even stronger result: The estimated standard error in each individual sample is a good guide to the actual variability of the OLS estimates across all samples.

[14] In fact, 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 estimates wouldn't converge to anything in particular. This means the standard deviation of the 2SLS standard error cannot be bootstrapped.

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, the red dots indicate $H_0 : \beta = 0$ is rejected at the 5% level using the 2SLS t-test, while 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%.

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. 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. We can start to understand the source of this phenomenon using exact finite sample theory. Phillips (1989) derives two key properties of 2SLS in the unidentified case (when the instrument is irrelevant). First, the 2SLS estimator converges in distribution to a scale mixture of normals centered on $E(\hat{\beta}_{OLS})$. Second, the 2SLS variance estimator ($\hat{\sigma}^2$) converges in distribution to a quadratic function of $\hat{\beta}_{2SLS}$, with a minimum at $E(\hat{\beta}_{OLS})$. This means 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 $\hat{\beta}_{2SLS}$. 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, the properties of 2SLS in the unidentified case have a major influence on 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. In Keane and Neal (2021) we give an intuitive explanation for why the association between 2SLS estimates and their standard errors exists in identified models, and we fully explore the implications of this phenomenon.[15] 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 much smaller standard errors.

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 hard 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 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 how the red shaded area is very small. In fact, the false null hypothesis is only rejected 5.1% of the time. This is an abysmally low level of power. In fact, if the null hypothesis were true, we would expect a well behaved 5% level t-test to reject it 5% of the time, and this is scarcely better than that!

In the right panel of Figure 1 we shade in red the cases where the 2SLS t-test rejects the null hypothesis that the true Frisch elasticity is equal to the true value of 0.597. The test rejects the null hypothesis 6.6% of the time. This is not so bad when viewed in isolation, as it is not too far from the correct rate of 5%. But more importantly, the rate of rejecting the true hypothesis that the Frisch equals 0.597 is actually ***greater*** than the rate of rejecting the false hypothesis that the Frisch equals 0. This is truly awful behavior for a statistical test.

[15] Basically, this pattern arises for the following reason: As we discussed in Section III, in the original NLS sample the partial correlation between the ASVAB score and wage growth is 0.04. When we look across our 10,000 subsamples, the correlation fluctuates around that value due to sampling variation. Two things happen in samples where that correlation is relatively high:

First, the 2SLS standard error estimate is smaller: The stronger is the correlation between the instrument and the endogenous variable, the smaller is the 2SLS standard error.

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).

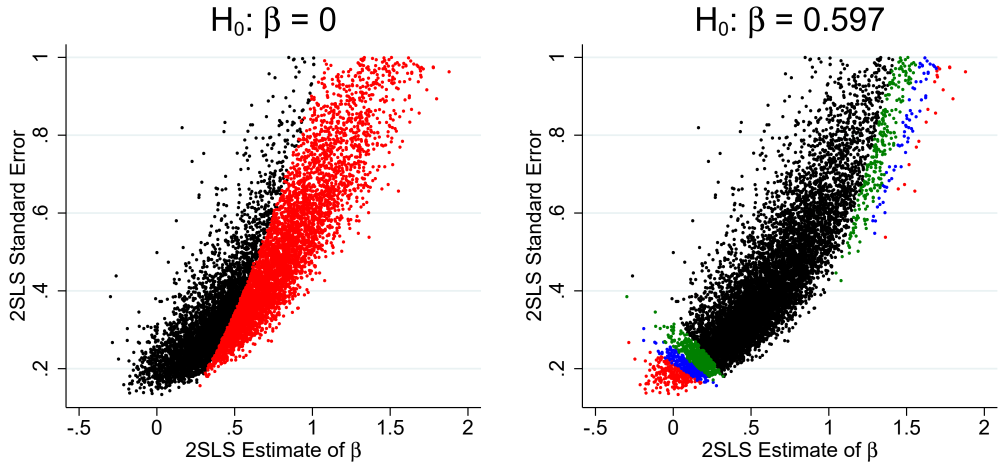
Putting these two facts together, it means that 2SLS estimates that are most shifted in the direction of the OLS bias (negative) appear to be more precise. This is exactly the pattern we see in Figure 1.

Another notable aspect of the right panel of Figure 1 is that the cases where we reject the null of Frisch = 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.

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 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 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 acheived 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 using the AR test. In the right panel red dots indicate* $H_0 : \beta = 0.597$ *rejected at the 5% level using the AR test. Blue and green indicate rejections at the 10% and 20% levels, respectively.*

The right panel of Figure 2 shows the rate of rejecting the true null hypothesis that the Frisch equals 0.597. The 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" to weak instruments, meaning that is has correct size (rejection rates) even if instruments are weak or borderline. Thus we see that *the AR test has correct size and ten times the power of the t-test.*

The only limitation of the AR test is that it doesn't quite generate symmetric rejections when the estimates are above and below the true value. For example, of the 4.9% rejections in the 5% test, 3.6% occur when the estimate is below 0.597 and 1.3% occur when it is above. This is because, like the t-test, the AR test tends to attribute greater precision to estimates shifted in the (negative) direction of the OLS bias, and less precision to large positive estimates. But this problem is much less severe for the AR test than the t-test.[16]

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. Yet, in the extensive literature on estimating the Frisch elasticity, we are not aware of any work that has used the AR 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. In Keane and Neal (2021) we show that the superiority of the AR test is evident across a wide range of contexts. 2SLS t-tests perform poorly in general due to the strong association between 2SLS estimates and their standard errors. As a result of this pattern, t-tests have difficulty detecting true negative (positive) effects when the OLS endogeneity bias is positive (negative). This problem is relevant across a wide range of empirical applications, including cases where instruments are much stronger than here. The AR test is much less susceptible to this problem.

VII. Interpreting the Empirical Results in Light of the Experiment

Returning to our empirical results in Table 1, we now assess them based on what we have learned from the Monte Carlo study. Recall that our 2SLS estimate of the Frisch elasticity based on the ASVAB instrument is 0.597, but the 2SLS t-test indicates this is not significantly different from zero at the 5% level. However, while the ASVAB score is a highly significant predictor of wage growth, it only explains a small fraction of the variance. The Monte Carlo experiment clearly indicates that the t-test is not reliable in this weak instrument environment.

Much more reliable is the weak instrument robust AR test, which is based on the significance of the instrument (ASVAB) in the reduced from regression of hours changes on wages changes. The AR test indicates that our Frisch elasticity estimate is significant at the 3.5% or 1.8% level, depending on whether we rely on the heteroskedasticity robust or cluster robust standard error.

[16] In Keane and Neal (2021) we show that the power asymmetry in the AR test vanishes quickly as instruments become stronger. But the power asymmetry in the t-test remains substantial even with very strong instruments.

We can also invert the AR test to obtain a weak instrument robust confidence interval, as discussed in Anderson and Rubin (1949).[17] 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.

VIII. Results Based on Multiple Instruments

As we discussed in Section II, much of the prior work on estimating the Frisch elasticity used education as the instrument for wage growth, but we rely on the ASVAB score as we find it is a stronger instrument in the first stage of 2SLS. In this section we consider using both education and the ASVB score as instruments. In order to keep the sample identical to that in Table 1, we code education as zero if it is missing, and introduce a dummy for missing education as an additional instrument. As we see in the first column of Table 3 both the ASVAB score and education are significant in the first stage of 2SLS, suggesting they capture somewhat different dimensions of ability.[18]

We also report two versions of the partial F-statistic for joint significance of the instruments in the first stage (heteroskedsaticity and cluster robust), as well as the Olea-Pfleuger effective F-test for weak instruments in a non-*iid* setting. These statistics range from 4.3 to 5.1, so they are well below conventional weak instrument testing thresholds. 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 1.017, which is much larger than the estimate of 0.597 we obtained in Table 1. Notably, the heteroskedasticity robust standard error increases from 0.403 to 0.481, so t=2.12 (p=.034) and a 5% t-test judges our estimate significant.[19] It may seem surprising that the 2SLS standard error increases despite the efficiency gain from adding an additional relevant instrument in the first-stage. But we have discussed how the increase in the Frisch estimate from .597 to 1.017, which moves us further from the OLS bias, will mechanically cause the 2SLS standard error of regression to increase. This tends to inflate the standard error of the 2SLS estimate.

Now 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. The cluster robust version of the AR test gives a p value of .0026. Moreover, the AR test is not the most powerful test in the over-identified case: the weak instrument robust conditional likelihood ratio (CLR) test of Moreira (2003) is more efficient. The cluster-robust CLR test has a p-value of .0012, so the evidence for a positive Frisch elasticity based on the robust statistics is very strong.

[17] 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.

[18] To be precise, the p-values for education are .047 or .071 based on the cluster robust or heteroskedasticity robust standard error, respectively.

[19] The cluster robust standard error increases from 0.363 to 0.442, giving t=2.30 (p=.021).

Table 3—Frisch Elasticity - Over-identified Models

	2SLS 1^{st} Stage	2SLS 2^{nd} Stage	Reduced Form	GMM-2S 2^{nd} Stage	GMM-CU 2^{nd} Stage
Dependent Variable	ΔW	ΔH	ΔH	ΔH	ΔH
Wage Change		1.017		0.896	1.310
		(0.481)		(0.474)	(0.548)
		[0.442]		[0.433]	[0.487]
ASVAB Ability Score	0.028		-0.007		
	(0.014)		(0.017)		
	[0.012]		[0.016]		
Education	0.002		0.006		
	(0.001)		(0.003)		
	[0.001]		[0.002]		
Education Missing	0.033		0.033		
	(0.034)		(0.044)		
	[0.035]		[0.043]		
F-Stat (Hetero-σ Robust)	4.31		4.21		
p-value	0.005		0.006		
F-Stat (Cluster Robust)	5.14		4.75		
p-value	0.002		0.003		
Olea-Pfleuger Effective F	4.57				
Exogeneity Test (AR or J)		2.77		3.19	3.00
p-value		0.428		0.203	0.224
R^2	0.008		0.013		

Note: 'GMM-2S' refers to the 2-step GMM, while '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. $N = 5,931$.

So here we see a milder version of the pattern in Table 1: The 2SLS t-test implies the Frisch elasticity estimate is (just) significant at the 5% level, while the weak instrument robust statistics (the AR and CLR tests) imply much higher levels of confidence. The relative 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.

A useful feature of the AR test is that we can evaluate it at $\hat{\beta}_{2SLS}$ rather than 0 to obtain a test of the 2SLS over-identifying restrictions. This is just an F-test for joint significance of the excluded instruments in a regression of the 2SLS residuals on all the instruments. Henceforth, we refer to these as the AR(0) and AR($\hat{\beta}_{2SLS}$) tests to distinguish the two. As we see in Table 3, the AR($\hat{\beta}_{2SLS}$) test statistic is 2.77. The test is distributed $\chi^2(3)$ so the p-value is .428. Thus we cannot reject the exogeneity of the instruments. This is important, as a failure of the over-identification test would invalidate the AR test, as it would suggest 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(0) and AR($\hat{\beta}_{2SLS}$) statistics should be evaluated in conjunction.

To assess the relative performance of the AR and t-test in the three instrument case, we ran a Monte Carlo analysis like that of Section VI, but using the 2SLS estimated model in Table 3 as the data generating process.[20] In terms of power, we find that a 5% t-test rejects the false null $H_0 : \beta=0$ at a 60.2% rate, compared to 88.4% for the AR test, and 94.7% for CLR. So the ranking is as expected.

If we invert the AR test (cluster robust F version) we obtain a 95% confidence interval for the Frisch elasticity of 0.241 to 4.336, while inverting the CLR test gives 0.269 to 4.461.[21] These intervals sit comfortably above zero, and cover the range of values typically used to calibrate macro models.

Finally, the last two columns of Table 3 report the two-step and continuously updated GMM results. These GMM estimates of the Frisch elasticity are 0.896 and 1.310 respectively. Notice how the increase in the point estimate to 1.310, moving it even further from the OLS bias, coincides with a further increase in the GMM-CU standard error to 0.548. The GMM estimates and standard errors have the same positive covariance as the 2SLS estimates and standard errors. Thus the GMM standard errors are also unreliable in this context.

However, Stock and Wright (2000) develop a weak instrument 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=20.47. The test is distributed $\chi^2(3)$ so the p-value is .0001 and the Frisch estimate is highly significant. Finally, we consider Hansen's test of over-identifying restrictions. As we see in Table 3 the J-test has $p > 0.20$, indicating we cannot reject the exogeneity of the instruments. This is important, as a failure of the J-test would invalidate the S test.

In summary, in the over-identified case the weak instrument robust AR and S tests indicate that the 2SLS and GMM estimates of the Frisch elasticity are highly significant. We caution that both tests may reject $H_0{:}\beta=0$ either because the null is false or because the instruments are endogenous. Hence, before relying on the AR(0) and S test results, it is important to verify, as we have here, that the AR($\hat{\beta}_{2SLS}$) and Hansen J-tests do not reject exogeneity of the instruments.[22] However, we emphasize that failure of the exogeneity tests would invalidate 2SLS t-test results as well, so the reliance of the AR(0) and S test results on validity of the instruments is not a disadvantage of these robust tests relative to the t-test.

[20] 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-identifed case the instruments do have small correlations with the 2SLS residuals. We need to partial out those correlations to set up the experiment.

[21] We use the Stata command developed by Finlay and Magnusson (2009) to implement the cluster robust version of the CLR test and to do the inversion.

[22] In the single endogenous variable, K instrument case, the AR($\hat{\beta}_{2SLS}$) 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.

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 is a vital input into tax policy. For example, higher values of the Frisch imply lower optimal tax rates on labor income. Because of its importance, there is a large literature estimating the Frisch elasticity using instrumental variable methods. But this literature has been plagued by weak instrument problems because it is hard to find instruments that strongly predict wage growth. Hence the value of the Frisch elasticity remains a topic of intense debate.

Here we revisit that debate. Using the ASVAB ability test as an instrument for wage growth, we estimate a 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 of the Frisch elasticity is very imprecise. Based on this, we can't even reject the hypothesis that it is zero at conventional levels – a result that is typical of many prior papers.

Importantly, the first stage F-statistic for our ASVAB instrument is 10.12, which is right on the borderline for whether weak instrument problems are a concern. This is again typical of prior work on estimating the Frisch elasticity.

Econometric theory strongly suggests that if weak instruments are a concern, 2SLS t-tests are unreliable, and the Anderson-Rubin (AR) test should be used instead. The AR test is robust to weak instruments and it is efficient. When we implement the AR test, we find our estimate of the Frisch elasticity is significantly greater than zero at the 5% level. In fact, it is significant at the 3.5% level.

These contradictory results led us to conduct an experiment to evaluate the reliability of the t-test vs. the AR test. In our data environment we find the AR test has correct size and ten times the power of the t-test. In fact, the power of the t-test is so poor that a 5% level test is more likely to reject a hypothesis that the Frisch equals its true value than a false hypothesis that it equals zero.

Given the clear theoretical guidance, along with empirical and Monte Carlo results like we present here, it is difficult to understand why applied researchers have not widely adopted the AR test in preference to 2SLS t-tests. The most likely explanation is that they are simply not aware of the severity of the problems with 2SLS t-tests that we document here.[23] When we use the appropriate inferential procedure (the AR test) we conclude the Frisch elasticity is fairly large (.597) and highly significant for young males in the NLSY97.

Our estimated Frisch elasticity for young men (.597) may still appear small compared to the values of 1.0 or more often used to calibrate macro models.

[23] In addition, applied researchers may be unfamiliar with AR tests and think they are difficult to implement. That is obviously not true, but the econometric theory literature on AR tests presents them at such a high level of generality that it is indeed difficult for applied researchers to penetrate. And of course, applied researchers may be wedded to t-tests simply because they are so familiar. But we hope that inertia may be overcome so that empirical practice can be improved.

However, there is accumulating evidence that the Frisch elasticity increases substantially with age (see, e.g., Borella, De Nardi and Yang 2019, Erosa, Fuster and Kambourov 2016, French 2005 and Keane 2021), and clear evidence that it is greater for women than men (see Keane 2011). So a value of .597 for young men is quite consistent with a value of 1.0 or more in the aggregate.

We also consider an over-identified model using both education and ASVAB as instruments for wage growth. Then our estimate of the Frisch increases to 1.02, and the 2SLS t-test indicates it is significant at the 3.4% level. However, the 2SLS standard error is again inflated due to its mechanical positive covariance with the estimate. An additional Monte Carlo experiment shows the AR test has correct size and twice the power of the t-test in this environment. The AR test gives a much higher significance level of 0.3%. If we invert the AR test we obtain a 95% confidence interval for the Frisch elasticity of 0.241 to 4.336, which covers the range of values typically used to calibrate macro models.

Of course, there have been numerous previous attempts to reconcile the small and sometimes insignificant 2SLS estimates of the Frisch elasticity that have often been obtained using micro data with the large values of the Frisch often used in macro calibrations. The various approaches are detailed in Keane and Rogerson (2012, 2015). These reconciliations fall into two broad categories: One set of explanations, exemplified by Imai and Keane (2004) and Domeij and Floden (2006) takes issue with the specification of equation (1), arguing that more general models of labor supply (e.g., models that account for human capital or liquidity constraints) imply that estimation of this equation will give downward biased estimates of the Frisch elasticity. 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. More recently, Gottlieb, Onken and Valladares-Esteban (2021) have shown how a large macro level Frisch elasticity can be reconciled with modest reactions to tax holidays due to a combination of income and equilibrium effects. These arguments are complementary to our argument here.

Our argument is new in that we criticise the micro-econometric literature on its own terms: Suppose the assumptions necessary for 2SLS estimation of equation (1) to deliver consistent estimates of the Frisch elasticity do 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 Frisch is both large and significant. We hope our straightforward econometric argument will prove convincing to economists who have not been convinced by the more subtle theoretical arguments based on more complex labor supply models or aggregation issues.

We conclude by re-stating our key econometric point: The main reason robust statistics lead to different conclusions about the Frisch elasticity from conventional 2SLS t-tests is a basic property of 2SLS that has been generally neglected: Specifically, when the OLS bias is negative, as is the case here, 2SLS estimates and their standard errors have a positive association (as they vary across random samples from the population). This mechanical positive association causes large positive estimates of the Frisch to appear spuriously imprecise, making it difficult for 2SLS t-tests to detect a true positive. Robust statistics like the Anderson and Rubin (1949) test are much less affected by this problem, so they are better able to detect a true positive Frisch elasticity.

In a different context, where the OLS bias is positive, this pattern would be reversed, and 2SLS standard errors on *positive* estimates would be spuriously precise. That would make 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.

In Keane and Neal (2021) we explore the implications of the association between 2SLS estimates and their standard errors in more detail. There we show that the serious problems with 2SLS t-tests that we have documented here persist even when instruments are strong, because the association between 2SLS estimates and their standard errors does not vanish as instrument strength increases. Thus it is advisable to use the AR test (or other robust tests) in lieu of the t-test even when instruments are strong.

ACKNOWLEDGEMENTS

We thank Peter Phillips for very helpful comments. This research was supported by Australian Research Council (ARC) grants DP210103319 and CE170100005. We have no conflicts of interest to declare.

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.

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.

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.

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

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.

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

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

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. 2021. "A Practical Guide to Weak Instruments." UNSW Economics Working Paper No. 2021-05a. Available at SSRN: https://ssrn.com/abstract=3846841.

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

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.