NBER WORKING PAPER SERIES

VALID T-RATIO INFERENCE FOR IV

David S. Lee
Justin McCrary
Marcelo J. Moreira
Jack R. Porter

Working Paper 29124
http://www.nber.org/papers/w29124

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2021, Revised March 2022

We are grateful to Charlie Fefferman for his generous spirit and interest in our problem, and to Peter Ozsváth for connecting us with him. We thank Isaiah Andrews, Josh Angrist, Esther Duflo, and Jim Stock for their comments and suggestions. We also thank Orley Ashenfelter, Marinho Bertanha, Stéphane Bonhomme, Janet Currie, Michal Kolesár, Alex Mas, José Montiel-Olea, Ulrich Mueller, Zhuan Pei, Mikkel Plagborg-Møller, Chris Sims, Eric Talley, Mark Wat-son, and participants of the joint Industrial Relations/Oskar Morgenstern Memorial Seminar at Princeton, the applied econometrics workshop at FGV, seminars at UC Davis and UQAM, the California Econometrics Conference, and the World Congress, for feedback on earlier iterations of this project. We are also grateful to Camilla Adams, Victoria Angelova, Cate Brock, Santiago Deambrosi, Colin Dunkley, Jared Grogan, Bailey Palmer, and Myera Rashid, and especially Sarah Frick and Katie Guyot for extraordinary research assistance. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Valid t-ratio Inference for IV
David S. Lee, Justin McCrary, Marcelo J. Moreira, and Jack R. Porter
NBER Working Paper No. 29124
August 2021, Revised March 2022
JEL No. C01,C1,C26,C36

ABSTRACT

In the single-IV model, researchers commonly rely on t-ratio-based inference, even though the literature has quantified its potentially severe large-sample distortions. Building on Stock and Yogo (2005), we introduce the tF critical value function, leading to a standard error adjustment that is a smooth function of the first-stage F-statistic. For one-quarter of specifications in 61 AER papers, corrected standard errors are at least 49 and 136 percent larger than conventional 2SLS standard errors at the 5-percent and 1-percent significance levels, respectively. tF confidence intervals have shorter expected length than those of Anderson and Rubin (1949), whenever both are bounded.

David S. Lee
Industrial Relations Section
Louis A. Simpson International Bldg.
Princeton University
Princeton, NJ 08544
and NBER
davidlee@princeton.edu

Justin McCrary
Columbia University
Jerome Greene Hall
Room 521
435 West 116th Street
New York, NY 10027
and NBER
jmccrary@law.columbia.edu

Marcelo J. Moreira
Department of Economics
Getulio Vargas Foundation - 11th floor
Praia de Botafogo 190
Rio de Janeiro - RJ 22250-040
moreira.marceloj@gmail.com

Jack R. Porter
University of Wisconsin-Madison
1180 Observatory Drive
6448 Social Sciences Building
Madison, WI 53706-1320
Madison, WI 53706-1320
jrporter@ssc.wisc.edu

For supplementary material, including updates to the original Online Appendix and a STATA package to compute tF critical values/standard error adjustments, please visit:
http://www.princeton.edu/~davidlee/wp/SupplementarytF.html

Consider the commonly employed single-variable, just-identified instrumental variable (IV) model, with outcome Y, regressor of interest X, and instrument Z,[1]

$$Y = \beta X + u, \text{ where} \tag{1}$$
$$C(u,Z) = 0, C(Z,X) \neq 0.$$

Conducting hypothesis tests and constructing confidence sets for β with correct significance and confidence levels has been pursued for several decades. In this setting, the validity of the Anderson-Rubin test (henceforth, *AR*) is well established (Anderson and Rubin, 1949)[2], and results expressing its advantages and optimality come in several flavors.[3]

Despite these findings, applied research, with rare exceptions, instead relies on t-ratio-based inference. Many studies have shown, numerically or theoretically, that the t-ratio test for IV significantly over-rejects and associated confidence intervals under-cover in situations when instruments are not sufficiently strong.[4] To deal with this problem, researchers have relied upon the first-stage F-statistic as a pre-test for instrument weakness. Staiger and Stock (1997) and Stock and Yogo (2005) provide a framework for precisely quantifying the distortions in—and therefore correcting—inference, with the use of the first-stage F-statistic. Importantly, although much of the econometric literature considers the general case of the over-identified model with multiple instruments, Stock and Yogo (2005) make clear that the distortions in inference also occur in the *single instrumental variable, just-*

[1]It will be shown that all of our results apply to the single excluded instrument case more generally, allowing for other covariates and variance estimators that accommodate departures from i.i.d. errors, such as heteroskedasticity-consistent, clustered, or time series approaches. Throughout, we use $V(\cdot)$ and $C(\cdot,\cdot)$ to denote population variance and covariance, respectively.

[2]Staiger and Stock (1997) show that *AR*-based inference delivers correct size/confidence with nonnormal and homoskedastic errors under arbitrarily weak instruments. Stock and Wright (2000), among others, show that *AR*-based inference is valid under more general error structures.

[3]The test of Anderson and Rubin (1949) in the just-identified case has been shown to minimize Type II error among various classes of alternative tests. These include classes of either unbiased tests (whose rejection probabilities under all alternatives are larger than that under the null) or invariant tests (which remain the same after transforming the data linearly). This is shown for homoskedastic errors, by Moreira (2002, 2009) and Andrews, Moreira and Stock (2006), and later generalized to cases for heteroskedastic, clustered, and/or autocorrelated errors, by Moreira and Moreira (2019).

[4]See, for example, Nelson and Startz (1990), Bound, Jaeger and Baker (1995), and Dufour (1997), and an earlier discussion by Rothenberg (1984). For a simple STATA program that demonstrates the inaccuracy of the standard approximation compared to the "weak-iv" asymptotic approximation, see http://www.princeton.edu/∼davidlee/wp/SupplementarytF.html

identified case—a common case for applied work, and the exclusive focus of the current paper.[5]

Unfortunately, the implementation and interpretation by practitioners of the approach and results of Staiger and Stock (1997) and Stock and Yogo (2005) has typically been imperfect or deficient. For example, pre-testing using the rule-of-thumb F-statistic threshold of 10 is commonplace, rather than the actual values provided in Stock and Yogo (2005) tables. Or, practitioners erroneously refer to the interval $\widehat{\beta} \pm 1.96 \cdot \widehat{\text{se}}\left(\hat{\beta}\right)$ as a "95% confidence interval" (after pre-testing using $F > 10$ as a diagnostic), even though the Bonferroni bounds of Staiger and Stock (1997) make clear that using $F > 16.38$ from Stock and Yogo (2005) implies that such an interval is in fact an 85% confidence interval.[6,7]

In the current paper, focusing on the single-instrument case, we meet practitioners "where they are" by introducing a new method of inference using only the first-stage F statistic and the 2SLS t-ratio. Rather than relying on a fixed pre-testing threshold value, we show how to smoothly adjust t-ratio inference based on the first-stage F statistic. In its simplest form, this amounts to applying an adjustment factor to 2SLS standard errors based on the first-stage F with the adjustment factors provided in tables below for 95% and 99% confidence levels. We refer to this procedure as the tF procedure and list some of its advantages here.

First, smooth adjustment yields usable finite confidence intervals for smaller values of the F statistic. In particular, for 95% confidence, finite adjustment factors are available for any value of $F > 3.84$. This puts the smooth adjustment approach on equal footing with AR, which yields bounded 95% confidence intervals for $F > 3.84$. Second, the confidence levels specified with the tF adjustment

[5]This single-variable case includes applications such as randomized trials with imperfect compliance (estimation of LATE, Imbens and Angrist (1994)), fuzzy regression discontinuity designs (see discussion in Lee and Lemieux (2010)), and fuzzy regression kink designs (see discussion in Card et al. (2015)).

[6]We write $\widehat{\beta}$ for the IV estimator and $\widehat{\text{se}}(\cdot)$ for the estimated standard error of an estimator.

[7]In their formulation, Staiger and Stock (1997) point out that this inferential statement requires a pre-commitment to a confidence set that is the *entire real line* in the event that F<16.38. Hall, Rudebusch and Wilcox (1996) show that over-rejection can be even worse in the presence of pre-testing for weak instruments. Andrews, Stock and Sun (2019) also discuss in detail the practice of selectively dropping specifications when first-stage F-statistics do not meet a particular threshold, and show that severe distortion can result.

factors leave little room for practitioner misinterpretation. These confidence levels incorporate the effects of basing inference on the first-stage F; again, this puts the confidence interval on equal footing with AR, or other procedures that have zero distortion. Third, even though the tF critical value function tends to infinity as F approaches 3.84 from above (e.g., for the 5 percent test), any alternative function that is uniformly below the tF critical value function in a neighborhood of 3.84 leads to over-rejection for some data generating process.

Fourth, our table of adjustment factors is "robust" to commonly considered error structures (e.g., heteroskedasticity or clustering). That is, no further adjustment is needed for these scenarios as long as the same type of robust variance estimator is used for the first-stage as for the IV estimate itself. Fifth, we compare the tF approach to AR based on expected confidence interval length. Given the well-established power properties of AR, our results here are surprising: conditional on $F > 3.84$, the expected length of the AR interval is *infinite*, while that of the tF interval is *finite*. Sixth, the tF adjustment can be easily applied to re-assess studies that have already been published, provided that the first-stage F-statistic has been reported, and does not require access to the original data.

In order to gauge the likely magnitude of tF adjustments in applied research going forward, we use a sample of studies recently published in the *American Economic Review* (AER) that utilize a single-instrument specification. For at least one-quarter of the specifications where the first-stage F-statistic is reported or can be computed from the published tables, applying the tF adjustment to the standard errors leads to an increase in confidence interval lengths of at least 49 and 136 percent for 5-percent and 1-percent significance levels, respectively. We observe that among the specifications for which $F > 10$ and $t^2 > 1.96^2$ (for the null hypothesis that the slope coefficient is zero)—which without our adjustment would likely have been deemed "statistically significant"—the use of tF adjustment would cause about one-fourth of the specifications to be statistically insignificant at the 5-percent level. We conclude therefore that these adjustments are likely to have a substantive impact on inferences in applied research that employ t-ratio inferences.

The paper is organized as follows. Section I uses recent papers published in the AER to characterize current inferential practices for the single-instrument IV

model. In Section II, we first describe the tF procedure—the critical values, the main results on power, and its application to our sample of studies. Section III describes how the results stated in Section II are derived. Section IV concludes.

I Inference for IV: Current Practice

To motivate our emphasis on improving t-ratio-based inference, this section documents facts about current practice for the single instrumental variable model, as reflected by recent research published in the *American Economic Review*. We later use this sample of studies to gauge to what extent our proposed adjustments could make a difference in practice.

Our sample frame consists of all *AER* papers published between 2013 and 2019, excluding proceedings papers and comments, yielding 757 articles, of which 123 include instrumental variable regressions. Of these 123 studies, 61 employ single instrumental variable (just-identified) regressions.[8] Consistent with the conclusion of Andrews, Stock and Sun (2019), this confirms that the just-identified case is an important and prevalent one, from an applied perspective.

From these papers, we transcribe the coefficients, standard errors, and other statistics associated with each *IV* regression specification. Each observation in our final dataset is a "specification," where a single specification is defined as a unique combination of 1) outcome, 2) endogenous regressor, 3) instrument, and 4) combination of covariates. The dataset contains 1311 specifications from 61 studies; among those studies, the average number of specifications is 21.5, with a median of 9, and with 25th and 75th percentiles of 4 and 21, respectively. The purpose of our dataset is to fully characterize specifications that are reported in published studies.[9]

Each specification is placed into one of four categories, as shown in Table 1, according to the types of regressions for which coefficients and standard errors

[8]Specifically, we include papers that exclusively employ just-identified specifications with one endogenous regressor and presented 2SLS results in the main text; i.e., we exclude a paper if it contains over-identified models, and we exclude papers if the only mention of a just-identified IV model is in an appendix.

[9]See Andrews, Stock and Sun (2019) for a more in-depth comparison of *AR* and t-ratio-based inference.

Table 1: Current Practice Implementing IV Estimation, Published Papers from AER

Combinations of regressions reported	First Stage F-statistic?		
	No	Yes	Total
Two-Stage Least Squares	445	132	577
	(0.339)	(0.101)	(0.44)
	[0.251]	[0.088]	[0.339]
Two-Stage Least Squares and First Stage	247	212	459
	(0.188)	(0.162)	(0.35)
	[0.204]	[0.154]	[0.358]
Two-Stage Least Squares and Reduced Form	13	7	20
	(0.01)	(0.005)	(0.015)
	[0.024]	[0.035]	[0.059]
Two-Stage Least Squares, First Stage, and Reduced Form	181	74	255
	(0.138)	(0.056)	(0.195)
	[0.15]	[0.094]	[0.244]
Total	886	425	1311
	(0.676)	(0.324)	(1)
	[0.628]	[0.372]	[1]

N=1311. Drawn from 61 published papers. Each observation represents a unique combination of outcome, regressor, instrument, and covariates. Unweighted proportions are in parentheses, and weighted proportions are in brackets, where the weights are proportional to the inverse of the number of specifications in the associated paper.

are reported: the coefficients and standard errors from 1) only the 2SLS, 2) the 2SLS and first-stage regression, 3) the 2SLS and the reduced-form regression of the outcome on the instrument, and 4) the 2SLS, the first-stage, and the reduced form. In addition, we identify whether or not, for each specification, the first-stage F-statistic is explicitly reported (as indicated by the first two columns in Table 1).[10]

For each configuration, Table 1 reports the number of specifications, proportions (in parentheses), and weighted proportions (in brackets) where the weight for each specification is the inverse of the total number of specifications reported from its

[10]We include in the second column F-statistics that were actually reported by authors as the "Kleibergen-Paap" (henceforth, KP) statistic from Kleibergen and Paap (2006), rather than as an F-statistic. As noted in Andrews, Stock and Sun (2019), in the case of a single endogenous-regressor with single instrument, $KP = F$. In our sample, about 39 percent (weighted) of the F statistics in the second column were reported as KP statistics.

study. Henceforth, unless otherwise specified, when we refer to proportions, we refer to the weighted proportions since we wish to implicitly give each study equal weight in the summary statistics that we report.

Table 1 shows that the most common combination among the eight possible types is the reporting of 2SLS coefficients without explicitly reporting the first-stage F-statistic, representing about a quarter of the specifications. The second most-common practice is to report both the 2SLS and the first-stage coefficients without reporting the F-statistic (about 20 percent), but it should be clear that the F-statistic can be derived from squaring the ratio of the first-stage coefficient to its associated estimated standard error. The least common reporting combination is 2SLS and the reduced form, without reporting the first-stage F (2.4 percent).

In our analysis of the data, in order to maximize the number of specifications for which we have a first-stage F-statistic, we compute it from the reported first-stage coefficients and standard errors, but whenever this is not possible, we use the explicitly reported F-statistic.[11]

Figure 1 displays the histogram of the F-statistics in our sample on a logarithmic scale. The weighted 25th percentile, median, and 75th percentiles are 14.23, 45.84, and 225, respectively. The figure shows that most of the reported first-stage F-statistics in these studies do pass commonly cited thresholds such as 10.[12] More detail on these specifications is provided in Table 2, which is a two-way frequency table for whether or not the square of the t-ratio for the hypothesis that $\beta = 0$ exceeds 1.96^2, and whether or not the computed F statistic exceeds 10 (a commonly-used or cited threshold). Overall, the table indicates that for about 60 percent of the specifications, the estimated 2SLS coefficient would be "statistically significant" under the practice of using a critical value of 1.96 and a first-stage F-statistic

[11] We find that among studies in which both the reported and computed F-statistic are available, about 63 percent of the time the two numbers are within 5 percent of one another. For those specifications in which the reported $\widehat{F}$ is the only F-statistic available, there are some situations where it is not entirely clear whether the F-statistic is the first-stage F; it is possible that they are F-statistics for testing other hypotheses.

[12] Consistent with the pattern observed in Andrews, Stock and Sun (2019), we observed in our sample that among those specifications where the F (or KP) statistics were explicitly reported, KP statistics were somewhat smaller: the weighted median KP statistic was 14.23, and among all the reported statistics below 10, about 61 percent were reported as KP statistics.

Table 2: t^2 and First-stage F-statistics, Conventional Critical Value, Rule of Thumb Threshold of 10

	F<10	F≥10	Total
$t^2 \geq 1.96^2$	64	408	472
	(0.076)	(0.482)	(0.557)
	[0.104]	[0.595]	[0.699]
$t^2 < 1.96^2$	41	334	375
	(0.048)	(0.394)	(0.443)
	[0.062]	[0.238]	[0.301]
Total	105	742	847
	(0.124)	(0.876)	(1)
	[0.167]	[0.833]	[1]

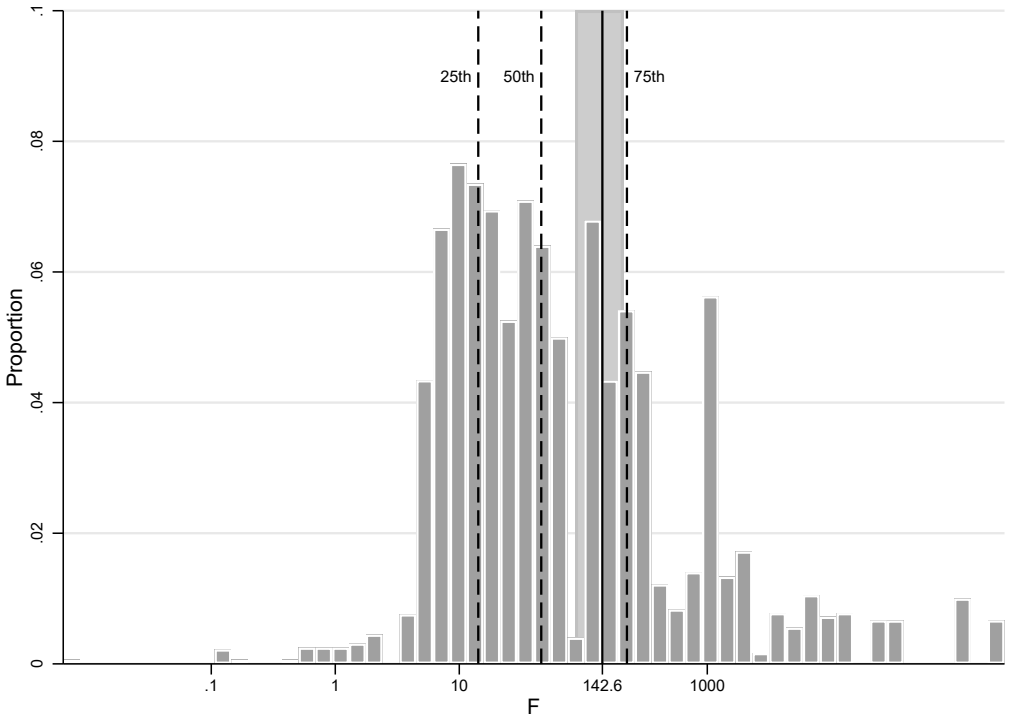
N=847. Unweighted proportions are in parentheses, and weighted proportions are in brackets. See notes to Table 1. All specifications use the derived F-statistic, and when not possible, the reported F-statistic. F-statistics can be derived for specifications that report nonzero standard errors in the first-stage; 6 specifications that report (rounded) first-stage standard errors of zero and do not report F-statistics are excluded.

threshold of 10 as a basis of trusting the inference.

We recognize that the null hypothesis of $\beta = 0$ may not always be the hypothesis of interest across all the studies. Furthermore, in our data collection, we do not make any judgments as to the extent to which any particular regression specification is important for the conclusions of the article. Indeed, in some cases, the 2SLS specification is used for a "placebo" analysis, where insignificant results are consistent with the identification strategy of the paper. In that spirit, it is beyond the scope of our paper to determine whether or not any particular study's overall conclusions are still supported despite any changes to the statistical inferences caused by using the corrections that we describe below. Instead, we focus more narrowly on gauging to what extent the tF critical values are likely to impact the length of confidence intervals in research going forward, using a recent sample of published studies to guide and inform that estimate.

Most importantly, we observe from our sample that *AR* test statistics or *AR* con-

Figure 1: Distribution of First-stage F-statistics



N=847 specifications. Scale is logarithmic. All specifications use the derived F-statistic or, when not possible, the reported F-statistic. F-statistics can be derived for specifications that report nonzero standard errors in the first-stage. Six specifications that report (rounded) first-stage standard errors of zero and do not report F-statistics are excluded. Proportions are weighted; see notes to Table 1. Dashed lines correspond to the 25^{th} (14.23), 50^{th} (45.84), and 75^{th} (225) percentiles of the distribution. The shaded region denotes the range between the 0.5^{th} and 99.5^{th} percentiles of a non-central χ^2_1 distribution with a non-centrality parameter equal to 142.6.

fidence regions are reported for less than 3 percent of the specifications, despite the fact that the econometric literature has noted that *AR* inference is valid and robust to weak instruments and has a number of other attractive properties; see the discussion, for example, in Andrews, Stock and Sun (2019). It is this stark difference between theoretical considerations and practice that motivates our focus. We surmise that practitioners may elect to use t-ratio inference, not because they believe it has superior properties compared to *AR*-based inference, but rather because it is presumed that any inferential approximation errors associated with the conventional t-ratio are minimal or acceptable. Or practitioners may presume that the inference has

the intended significance or confidence level, as long as the observed first-stage F-statistic is sufficiently large—even though Stock and Yogo (2005) explicitly point out that using 1.96 critical values can lead to over-rejection (or under-coverage) even *with* the use of their critical values for the F-statistic.

tF inference eliminates this known and quantified distortion, taking as given the common practice of computing the 2*SLS* and standard errors and providing critical values that result in the intended significance or confidence levels.

An additional and separate motivation for exploring alternatives to *AR* is that, if our sample is any indication, there are likely hundreds of other published studies that use the single-instrument *IV* model, most of which do not use *AR*-based inference. In many cases, it may be prohibitively costly to obtain the original data to assess how inferences might change when using *AR*. The adjustment we introduce below allows one to adjust the reported 2SLS standard error solely on the basis of the already-reported (or implicitly computed) first-stage F-statistic.

II Valid t-based Inference: Theoretical Results and Empirical Implications

This section states our main theoretical findings, emphasizing the motivation for the tF procedure, and how to use the critical value tables in practice. We defer the derivations of our results to Section III, and details of the proofs to the Online Appendix.

We begin by briefly reviewing the inferential problem with the t-ratio for *IV*, as already established in the econometric literature. This motivates tF as a solution to that problem. We then present the tF critical values for the 5 percent and 1 percent levels.[13] Since the use of the tF critical values allows one to achieve intended significance and confidence levels, we then present some results on how the power of the tF procedure compares to that of *AR*. Finally, we describe how the applica-

[13]We focus the specific cases of obtaining valid tests at the 5 percent and 1 percent significance levels and the corresponding 95 percent and 99 percent confidence intervals, because these standards of evidence are commonly used in applied research. However, it will be clear in Section III that our formulas can be adapted to analyze other levels of significance or confidence levels.

tion of the tF adjustments impacts the statistical inferences in our sample of *AER* studies.

II.A The tF procedure: Notation and Motivation

We begin with the notation for the structural and first-stage equations including additional covariates:

$$Y = X\beta + W\gamma + u$$
$$X = Z\pi + W\xi + v$$

where W denotes the additional covariates which can include a constant corresponding to an intercept term. Without loss of generality, we assume orthogonality between W and each of Y, X, Z.[14]

The key statistics are given by

$$\hat{t} \equiv \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{V}_N\left(\hat{\beta}\right)}} \quad \text{and} \quad \hat{f} \equiv \frac{\hat{\pi}}{\sqrt{\hat{V}_N\left(\hat{\pi}\right)}},\ \hat{F} = \hat{f}^2$$

where $\hat{\beta}$ is the instrumental variable estimator. $\hat{V}_N\left(\hat{\beta}\right)$ represents the estimated variance of $\hat{\beta}$, which can be a consistent robust variance estimator to deal with departures from i.i.d. errors, including one- or two-way clustering (e.g., see Cameron, Gelbach and Miller (2011)). $\hat{t}$ is the usual t-ratio, where we first consider the distribution of this statistic when the null hypothesis is true, but later on, when discussing power in greater detail, we make the distinction between the true value β and the (possibly false) hypothesized value β_0. $\hat{f}$ is the t-ratio (for the null hypothesis that $\pi = 0$) for the first-stage coefficient, and its square is equal to the F-statistic, which we denote $\hat{F}$.

[14] All of our results allow for covariates, since one can redefine Y, X, and Z as the residual from regressing each of those variables on W. Using these residuals after partialing out the covariates delivers the exact same point estimates, and standard errors, as if 2SLS was employed including the covariates.

The traditional argument for t-ratio inference is as follows. Under the null hypothesis $\hat{t}^2 \xrightarrow{d} t^2$. That is, the argument is that in large samples, a good approximation of the statistic $\hat{t}$ is the random variable t, a standard normal, with its square therefore being a chi-square with one degree of freedom. This approximation underlies the use of the standard normal critical values ± 1.96 for testing hypotheses at the 5 percent level. More generally, the critical values $\pm\sqrt{q_{1-\alpha}}$ are used for tests at the α level of significance, where $q_{1-\alpha}$ is the $(1-\alpha)$th quantile of the chi-squared distribution with one degree of freedom.

What has been established and understood in the theoretical literature for quite some time—but perhaps not fully internalized by practitioners more broadly—is that 1) the use of a standard normal to describe the distribution of the random variable t can lead to systematically distorted inference even with very large samples, and 2) the magnitude of the distortion can be precisely quantified. More specifically, it has been understood in the econometric literature that even when samples are large, t has a known *non-normal* distribution, which in some cases might be "close" to the standard normal, but in other cases, the deviation from normality can be significant.

Specifically, Stock and Yogo (2005) derive a formula for using Wald test statistics based on 2SLS (and other k-class estimators). In the just-identified case with one endogenous regressor, their results show that t^2 under the null can be seen as a function of two jointly normal random variables. With some re-arrangement of terms, the two normal variables can be seen as f and t_{AR}, where $\hat{f} \xrightarrow{d} f$ and f has mean $f_0 \equiv \frac{\pi}{\sqrt{\frac{1}{N}AV(\hat{\pi})}}$ and unit variance, where $AV(\hat{\pi})$ is the asymptotic variance of $\hat{\pi}$ and t_{AR} is a standard normal with $AR = t_{AR}^2$. The correlation ρ of f and t_{AR} is the correlation of Zu and Zv.[15]

Their t^2 formula allows one to precisely quantify the degree of distortions in inference from using the rule $t^2 > q_{1-\alpha}$ to reject the null hypothesis. Based on this formula, Panel (a) of Figure 2 provides a visualization of this relationship: it graphs rejection probabilities—the probability that $t^2 > 1.96^2$ under the null hypothesis—

[15] When the data are homoskedastic, ρ simplifies to the correlation between u and v. Stock and Yogo (2005) use a homoskedastic model.

for different values of $E[F]$ and ρ, where $E[F] = f_0^2 + 1$.[16] The figure illustrates that with low values of ρ (e.g., 0 or 0.5)—a lower degree of "endogeneity"—the t-ratio rejects at a probability below the nominal 0.05 rate. On the other hand, for $\rho = 0.8$, for example, the rejection rate can be as large as 0.13, when the instrument is close to irrelevant. In the extreme, with a maximal value of ρ equal to 1, the rejection probability tends to 1 as instruments become arbitrarily weak. The true significance level (size) of any test is by definition the maximum rejection probability across all possible values of the nuisance parameters – here, ρ and $E[F]$. Thus, the test based on $t^2 > q_{1-\alpha}$ clearly has incorrect size, as widely understood in the econometric literature. Indeed Stock and Yogo (2005) explicitly provide the quantity represented by the red circle in Figure 2 Panel (a): when $\rho = 1$ and $E[F] = 6.88$, the rejection probability is 0.10; it represents the minimum value of $E[F]$ one needs to assume in order for the ± 1.96 critical values will lead to significance level of 0.10.

Even though one does not know the values of ρ or $E[F]$, Staiger and Stock (1997) and Stock and Yogo (2005) propose to use the observed first-stage $\hat{F}$. Re-expressing the t^2 formula in Stock and Yogo (2005) in terms of f and t_{AR}, as mentioned above, we can determine

pairs of critical values c^* and F^* such that

$$\Pr\left[t^2 > c^*, F > F^*\right] \leq \alpha$$

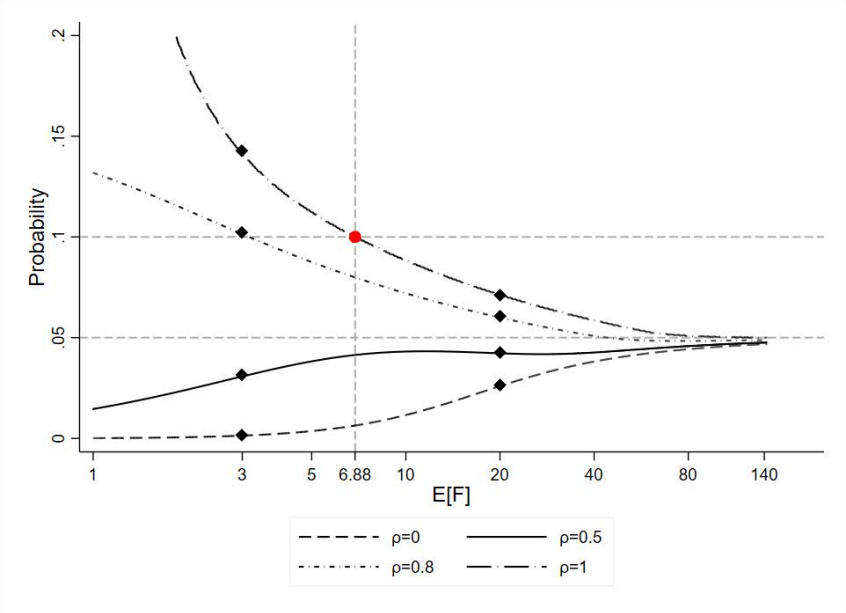
for a pre-specified significance level α. This amounts to a "step function" critical value function: if $F < F^*$, set $c^* = \infty$ (accept the null); otherwise, use the value c^* as the critical value for t^2.[17] Put equivalently, this implies a confidence interval procedure that sets the confidence interval to the entire real line if $F < F^*$, and otherwise uses $\pm\sqrt{c^*} \times$ *(standard error)* for the confidence interval.

[16] As we explain in detail in Section III, rejection probabilities displayed in Figure 2 Panel (a) can be computed directly from integral expressions, and are accurate up to the precision of numerical integration. To provide assurance that our formulas and numerical integration give correct answers, we additionally performed Monte Carlo simulations, and we plot examples of those results as diamonds in Figure 2. Those results match quite closely with our theoretical calculations.

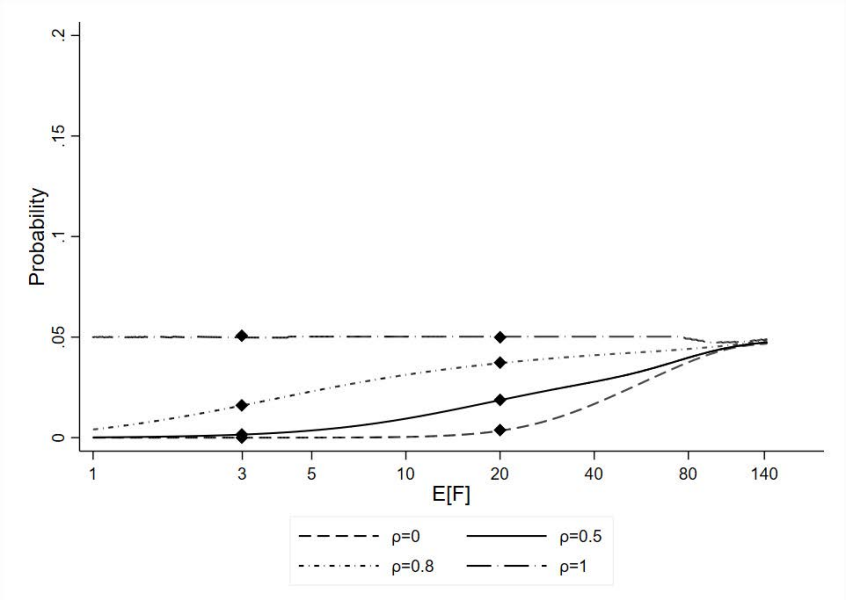
[17] This approach is in the same spirit as the Bonferroni confidence regions discussed in Section 4B of Staiger and Stock (1997). Using their approach, captured by their Equation (4.2), one can use $F^* = 16.38$ (as reported in Stock and Yogo (2005)) and $c^* = 1.96^2$ to obtain intervals with 85 percent confidence, while remaining agnostic about the true strength of the first stage.

Figure 2: Rejection Probabilities for t^2 and tF

(a) $\Pr[t^2 > 1.96^2]$ vs. $E[F]$, for selected values of ρ



(b) $\Pr[t^2 > c_{0.05}(F)]$ vs. $E[F]$, for selected values of ρ



Note: The x-axis scale is $\ln(E[F])$. The red circle in panel (a) corresponds to the quantity reported in Stock and Yogo (2005). A black diamond represents the rejection probability from 250,000 Monte Carlo simulations, each with a sample size of 1,000.

Utilizing the same analytical expressions in Stock and Yogo (2005), this paper introduces the tF critical value function $c_\alpha(F)$ such that

$$\Pr\left[t^2 > c_\alpha(F)\right] \leq \alpha$$

for a pre-specified significance level α, where $c_\alpha(F)$ is a smooth function of F, instead of a step function.[18] As we will show below, inference based on tF has significant power advantages over inference based on a test that uses constant thresholds c^*, F^*; furthermore, tF confidence intervals will have shorter expected length compared to that of AR when both are bounded intervals.

II.B The tF procedure: critical values and valid inference

Table 3 Panel A reports numbers that reflect the shape of the function $c_{0.05}(F)$. Specifically, corresponding to each value of the first-stage F-statistic (the first line of numbers in each row), is the corresponding critical value $\sqrt{c_{0.05}(F)}$ for $|t|$ (the second line of numbers in each row). $\sqrt{c_{0.05}(F)}$ tends to infinity as F tends to 1.96^2 from above, and it is strictly decreasing in F until reaching a minimum, the constant value of 1.96, when F reaches around 104.7.

The third line of numbers in each row normalize the critical values by 1.96, and therefore represent a standard error adjustment factor. Adjusted standard errors can be constructed using the table as follows: 1) Estimate the usual 2SLS (e.g., robust, clustered, etc.) standard error, 2) multiply the standard error by the adjustment factor (third line of numbers in each row) in the table corresponding to the observed first-stage $\hat{F}$ statistic. This adjusted standard error should be called a "0.05 tF standard error", and can be used for constructing the t-ratio for testing a particular hypothesis, or for constructing 95% confidence intervals using $\hat{\beta} \pm 1.96 \times$ ("0.05 tF standard error"). Since the table contains selected values from an underlying convex function, to compute intermediate values, a conservative approach would be to linearly interpolate between the selected values. As an example

[18] Similar in spirit to the Bonferroni approach discussed in Section 4B of Staiger and Stock (1997), the probability considered is an unconditional one. See Chioda and Jansson (2005) for an analysis of inference conditional on the observed F-statistic.

of this interpolation, if the first-stage $\hat{F}$ is 10, one would multiply the estimated standard error by $1.727 + \frac{10.253-10}{10.253-9.835} \times (1.767 - 1.727) = 1.751$ to obtain the "0.05 tF standard error".[19]

It is important to note that these "adjusted standard errors" are valid only for 0.05 significance or 0.95 confidence levels. Different adjustments are needed for different significance/confidence levels. We report the analogous critical values and adjustment factors for corresponding selected values of F, for significance (confidence) levels of 0.01 (0.99), another commonly-used standard in applied research, in Table 3 Panel B.

The table shows that the $\frac{\sqrt{c_{0.01}(F)}}{2.576}$ function has a similar pattern, but three important differences. First, the adjustment factor now has a vertical asymptote at $F = q_{0.99} = 2.576^2$. Second, $c_{0.01}(F)$ declines until $F = 252.34$, at which point the adjustment factor is 1.059. Finally, we note that $\frac{\sqrt{c_{0.01}(F)}}{2.576}$ is uniformly strictly above $\frac{\sqrt{c_{0.05}(F)}}{1.96}$. This implies that from a reporting convenience standpoint, one could choose to report only the "0.01 tF standard errors" by using the adjustments in Table 3 Panel B, and the intervals $\hat{\beta} \pm 2.576 \times$ (".01 tF standard error") and $\hat{\beta} \pm 1.96 \times$ ("0.01 tF standard error") would be assured of confidence levels at both the 99th and 95th percent levels. The cost for this reporting convenience is that the latter interval would be unnecessarily conservative.

[19] We have also posted code at http://www.princeton.edu/~davidlee/wp/SupplementarytF.html to allow more precise computation of the adjustment factor for any given value of $\hat{F}$.

Table 3 Panel A: Selected Values of tF Critical Values, $\sqrt{c_{0.05}(F)}$, and tF Standard Error Adjustments, $\sqrt{c_{0.05}(F)}/1.96$

F	4.000	4.008	4.015	4.023	4.031	4.040	4.049	4.059	4.068	4.079
$\sqrt{c_{.05}(F)}$	18.656	18.236	17.826	17.425	17.033	16.649	16.275	15.909	15.551	15.201
$\sqrt{c_{.05}(F)}$ / 1.96	9.519	9.305	9.095	8.891	8.691	8.495	8.304	8.117	7.934	7.756
	4.090	4.101	4.113	4.125	4.138	4.151	4.166	4.180	4.196	4.212
	14.859	14.524	14.197	13.878	13.566	13.260	12.962	12.670	12.385	12.107
	7.581	7.411	7.244	7.081	6.922	6.766	6.614	6.465	6.319	6.177
	4.229	4.247	4.265	4.285	4.305	4.326	4.349	4.372	4.396	4.422
	11.834	11.568	11.308	11.053	10.804	10.561	10.324	10.091	9.864	9.642
	6.038	5.902	5.770	5.640	5.513	5.389	5.268	5.149	5.033	4.920
	4.449	4.477	4.507	4.538	4.570	4.604	4.640	4.678	4.717	4.759
	9.425	9.213	9.006	8.803	8.605	8.412	8.222	8.037	7.856	7.680
	4.809	4.701	4.595	4.492	4.391	4.292	4.195	4.101	4.009	3.919
	4.803	4.849	4.897	4.948	5.002	5.059	5.119	5.182	5.248	5.319
	7.507	7.338	7.173	7.011	6.854	6.699	6.549	6.401	6.257	6.117
	3.830	3.744	3.660	3.578	3.497	3.418	3.341	3.266	3.193	3.121
	5.393	5.472	5.556	5.644	5.738	5.838	5.944	6.056	6.176	6.304
	5.979	5.844	5.713	5.584	5.459	5.336	5.216	5.098	4.984	4.872
	3.051	2.982	2.915	2.849	2.785	2.723	2.661	2.602	2.543	2.486
	6.440	6.585	6.741	6.907	7.085	7.276	7.482	7.702	7.940	8.196
	4.762	4.655	4.550	4.448	4.348	4.250	4.154	4.061	3.969	3.880
	2.430	2.375	2.322	2.270	2.218	2.169	2.120	2.072	2.025	1.980
	8.473	8.773	9.098	9.451	9.835	10.253	10.711	11.214	11.766	12.374
	3.793	3.707	3.624	3.542	3.463	3.385	3.309	3.234	3.161	3.090
	1.935	1.892	1.849	1.808	1.767	1.727	1.688	1.650	1.613	1.577
	13.048	13.796	14.631	15.566	16.618	17.810	19.167	20.721	22.516	24.605
	3.021	2.953	2.886	2.821	2.758	2.696	2.635	2.576	2.518	2.461
	1.542	1.507	1.473	1.440	1.407	1.376	1.345	1.315	1.285	1.256
	27.058	29.967	33.457	37.699	42.930	49.495	57.902	68.930	83.823	104.67
	2.406	2.352	2.299	2.247	2.197	2.147	2.099	2.052	2.006	1.96
	1.228	1.200	1.173	1.147	1.121	1.096	1.071	1.047	1.024	1.00

The top number in each of the ten rows is the first-stage F statistic, the middle number is the corresponding critical value, $\sqrt{c_{0.05}(F)}$, and the bottom number in each row is the corresponding value of $\sqrt{c_{0.05}(F)}/1.96$, where we write 1.96 as a shorthand for $\Phi^{-1}(0.975)$. Numerical values in each pair are rounded up (e.g., 4.0051 rounds up to 4.006).

We verify that the tF adjustment achieves the intended significance level of 5 percent in Panel (B) of Figure 2, which is analogous to Panel (A), plotting rejection probabilities for the tF procedure for the same values of ρ and f_0. The curves are accurate up to the precision of our numerical integration. To provide some additional assurance that our formulas and numerical computations are correct, as in Panel (A), the diamonds represent Monte Carlo simulation rejection rates, which line up with the curves, as expected from the theory.

Table 3 Panel B: Selected Values of tF Critical Values, $\sqrt{c_{0.01}(F)}$, and tF Standard Error Adjustments, $\sqrt{c_{0.01}(F)}/2.576$

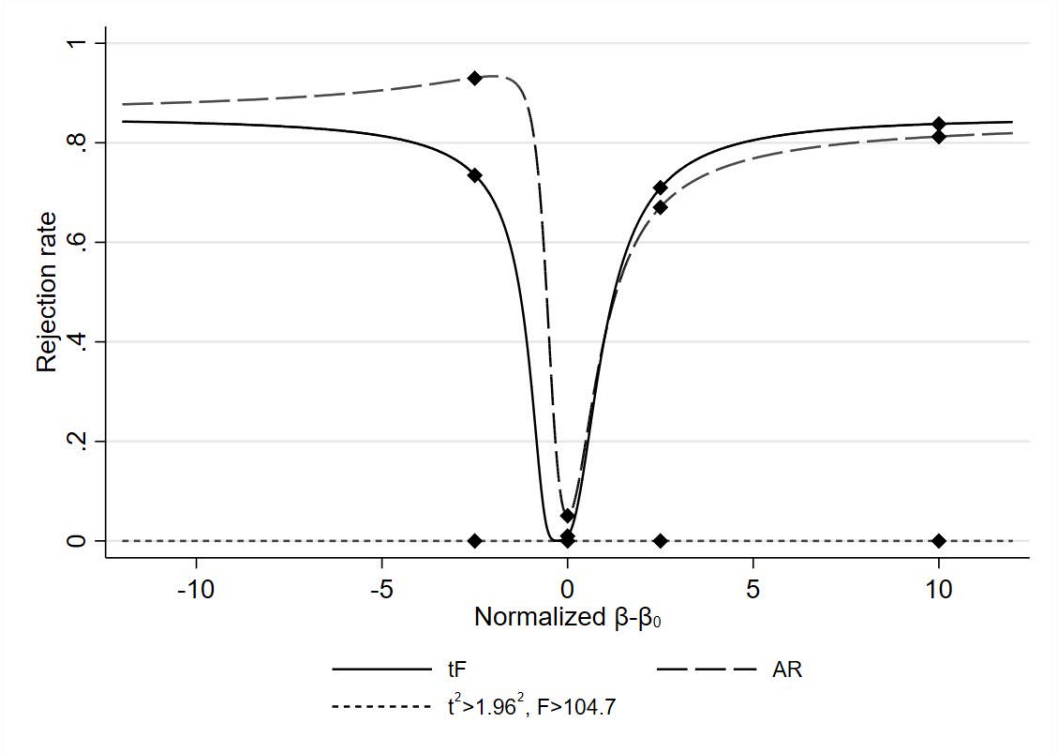
F	6.670	6.673	6.676	6.679	6.682	6.685	6.689	6.693	6.697	6.701
$\sqrt{c_{.01}(F)}$	91.097	87.924	84.862	81.907	79.054	76.301	73.644	71.079	68.604	66.214
$\sqrt{c_{.01}(F)}$ / 2.576	35.366	34.135	32.946	31.798	30.691	29.622	28.591	27.595	26.634	25.706
	6.706	6.711	6.717	6.723	6.729	6.736	6.743	6.751	6.759	6.768
	63.908	61.683	59.535	57.461	55.460	53.529	51.664	49.865	48.129	46.453
	24.811	23.947	23.113	22.308	21.531	20.781	20.058	19.359	18.685	18.034
	6.778	6.788	6.799	6.811	6.824	6.837	6.852	6.867	6.884	6.901
	44.835	43.273	41.766	40.312	38.908	37.553	36.245	34.983	33.765	32.589
	17.406	16.800	16.215	15.650	15.105	14.579	14.072	13.581	13.109	12.652
	6.920	6.941	6.963	6.986	7.011	7.038	7.066	7.097	7.129	7.164
	31.454	30.358	29.301	28.281	27.296	26.345	25.428	24.542	23.687	22.863
	12.211	11.786	11.376	10.980	10.597	10.228	9.872	9.528	9.196	8.876
	7.202	7.242	7.285	7.331	7.380	7.432	7.489	7.549	7.614	7.683
	22.066	21.298	20.556	19.840	19.149	18.482	17.839	17.218	16.618	16.039
	8.567	8.269	7.981	7.703	7.435	7.176	6.926	6.685	6.452	6.227
	7.757	7.836	7.922	8.013	8.111	8.216	8.329	8.451	8.581	8.721
	15.481	14.942	14.421	13.919	13.434	12.966	12.515	12.079	11.658	11.252
	6.010	5.801	5.599	5.404	5.216	5.034	4.859	4.690	4.526	4.369
	8.872	9.035	9.210	9.399	9.603	9.824	10.062	10.320	10.600	10.904
	10.860	10.482	10.117	9.765	9.425	9.097	8.780	8.474	8.179	7.894
	4.217	4.070	3.928	3.791	3.659	3.532	3.409	3.290	3.176	3.065
	11.235	11.595	11.988	12.418	12.889	13.407	13.979	14.610	15.312	16.094
	7.619	7.354	7.098	6.851	6.612	6.382	6.160	5.945	5.738	5.538
	2.958	2.855	2.756	2.660	2.567	2.478	2.392	2.308	2.228	2.150
	16.969	17.953	19.067	20.333	21.783	23.455	25.399	27.680	30.383	33.624
	5.345	5.159	4.980	4.806	4.639	4.477	4.321	4.171	4.026	3.885
	2.076	2.003	1.934	1.866	1.801	1.739	1.678	1.620	1.563	1.509
	37.560	42.416	48.511	56.324	66.592	80.502	100.069	128.950	174.370	252.342
	3.750	3.620	3.494	3.372	3.254	3.141	3.032	2.926	2.824	2.726
	1.456	1.406	1.357	1.309	1.264	1.220	1.177	1.136	1.097	1.059

The top number in each of the ten rows is the first-stage F statistic, the middle number is the corresponding critical value, $\sqrt{c_{0.01}(F)}$, and the bottom number in each row is the corresponding value of $\sqrt{c_{0.01}(F)}/2.576$, where we write 2.576 as a shorthand for $\Phi^{-1}(0.995)$. Numerical values in each pair are rounded up (e.g., 6.6712 rounds up to 6.672).

II.C The tF procedure: power comparisons to AR and step rules

In this subsection, we state our results on power, deferring derivations, proofs, and further discussion to Section III and the Online Appendix. Since the tF and AR tests (as well as rules like $t^2 > c^*, F > F^*$ with appropriately chosen c^* and F^*) can deliver inferences at the same intended significance/confidence levels under the same asymptotic approximation, it is natural then to investigate the relative power of these test procedures. For the purposes of this power comparison, we set $c^* = 1.96^2$ and use the minimum F^*—104.7—needed to ensure a test with significance level

Figure 3: Power curves for $\rho = 0.5$ and $f_0 = 3$



Note: A black diamond represents the rejection probability, from 250,000 Monte Carlo simulations, each with a sample size of 1,000.

0.05. We summarize the results below. Note that in our comparisons, we focus only on procedures that allow the researcher to be completely agnostic about the nuisance parameters.[20]

We produce standard power curves by generalizing the analytical expressions for the probability of rejection to depend on an additional parameter—a normalized deviation $\beta - \beta_0$, where β is the true parameter, while β_0 is the hypothesized value.[21]. We then compute the rejection probabilities with respect to this quantity for different scenarios according to the combination of nuisance parameters, ρ and f_0. Any combination of ρ and f_0 could be investigated: we illustrate these

[20]For example, the approach of Kocherlakota (2020) requires the researcher to assume a lower bound for f_0 for inference and thus is not among the approaches we consider.

[21]Specifically, the normalized $\beta - \beta_0$ is the unnormalized $\beta - \beta_0$ divided by $\frac{\sqrt{E[Z^2u^2]}}{\sqrt{E[Z^2v^2]}}$

traditional power curves for the nine combinations given by the three values of $\rho = 0, 0.5, 1$ and the three values of $f_0 = 1, 3, 9$.[22]

Figure 3 plots the power curve under the scenario $\rho = 0.5, f_0 = 3$ (which corresponds to $E[F] = 10$). It shows that tF and AR have roughly similar power, but neither uniformly dominates the other.[23] In particular, when the alternative value of β is sufficiently larger than β_0, then tF becomes slightly more powerful, while the opposite is true when β is smaller than β_0. An example of what this means for practitioners is that if the null is $\beta_0 = 0$, and $\rho > 0$ (which would imply that the OLS estimand is upward biased when errors are homoskedastic), then the probability of rejecting that null will be slightly higher for tF than for AR if the true effect is sufficiently positive.[24] Both tF and AR have a substantial power advantage over the step rule $c^* = 1.96^2, F^* = 104.7$. This latter observation should not be surprising since, in the scenario that $E[F] = 10$, the probability that F would exceed $F^* = 104.7$ is extremely low.

Appendix Figure A2 in Appendix A.9 includes power curves for the other eight scenarios for ρ, f_0. The pattern of results mirror those described above, with the additional observations that 1) the power curves for AR are consistently higher for $\rho = 0$, and 2) the differences between tF and AR (for any ρ) are negligible with $f_0 = 9$, but 3) the dependence of the relative power between tF and AR on the sign of $\beta - \beta_0$ remains apparent with high endogeneity ($\rho = 1$). The threshold rule continues to have low power in the nine scenarios we consider, which is not surprising since, even with $E[F] = 9^2 + 1 = 82$, the probability that F exceeds 104.7 continues to be relatively low. As f_0 increases so that the instrument is much stronger, the power curves for the step rule, tF, and AR all become closer to one

[22]To provide additional assurance in our theoretical derivations and implementation of numerical integration was carried out correctly, we overlay (as the diamonds in each graph) the results from Monte Carlo simulations, where we generate the underlying data according to each scenario and selected values of $\beta - \beta_0$ and compute the fraction of the time, over 250,000 Monte Carlo draws of sample sizes of 1,000 each, that each of the tests reject the null hypothesis. All of the results line up well with the theoretical values as computed from our analytical expressions for rejection probabilities.

[23]Note that while AR has known power optimality among unbiased tests, tF is not unbiased. The degree of bias can be seen in the power graphs.

[24]Note that the power curves are symmetric with respect to ρ; that is, when $\rho = -0.5$ then the power curve looks identical except the x-axis would be labeled $\beta_0 - \beta$.

another.

Given that neither AR nor tF uniformly dominates the other across all values of $\beta - \beta_0$ for fixed values of the nuisance parameters, we turn to a different and intuitive summary measure of power: the expected length of the confidence intervals for AR and tF conditional on $F > q_{1-\alpha}$. The reason why we focus on the condition $F > q_{1-\alpha}$ is that it is a necessary and sufficient condition for both the tF and AR confidence sets to be bounded intervals; when $F < q_{1-\alpha}$, both the AR and tF confidence sets are unbounded (i.e. have infinite length). The nonzero probability that $F < q_{1-\alpha}$ implies that the tF and AR confidence sets will have infinite *unconditional* expected length. Conditional on the event $F > 1.96^2$, it is immediately clear that the step rule of $c^* = 1.96^2, F^* = 104.7$ will also have infinite expectation since $104.7 > 1.96^2$.[25]

For any realization of the data, the tF and AR confidence sets behave similarly in the following sense: either both are bounded intervals (this happens when $F > q_{1-\alpha}$) or both are unbounded (this happens when $F \leq q_{1-\alpha}$). Thus, to compare expected lengths, we compare only the realizations of data that yield bounded intervals for both methods. That is, we compute expected conditional lengths conditional on $F > q_{1-\alpha}$. Surprisingly, our theoretical investigation reveals that the conditional expected length of the AR confidence interval is *infinite*. We show, by contrast, the conditional expected length for the tF interval is finite. We show below that this is true uniformly across all possible values of the nuisance parameters. This has a very straightforward implication for practitioners. Conditional on the event that they produce bounded intervals (which occurs with identical probabilities), the expected length of the tF confidence interval will always be shorter than the expected length of AR confidence intervals.

These findings are more fully described in Section III and proven in the Appendices C.2 and C.3. Here, we provide a simple visual of this result via a Monte Carlo exercise, shown in Figure 4.[26] Using the same data generating process from

[25] Indeed, Gleser and Hwang (1987) and Dufour (1997) show that in models which allow for non- (or nearly non-) identification, such as the IV model, any inference procedure with correct coverage *must* have infinite unconditional expected length.

[26] We use the Monte Carlo design from the discussion on single-variable *IV* in Angrist and Pischke (2009*a*), and discussed in Angrist and Pischke (2009*b*).

Figure 3, we run repeated Monte Carlo simulations of sample size 1,000 each. For each draw, we keep only those draws such that $\hat{F} > 1.96^2$, and when this occurs we compute the length of the *AR* and the tF confidence interval. For each specified number of Monte Carlo draws, we compute this conditional average using all accumulated draws up to that point. We do this four times, using an independent set of draws each time. The figure exhibits the patterns that one would expect to see if the conditional expected length were infinite for *AR* and finite for tF intervals: even after 500,000 draws, the conditional averages for *AR* do not appear to be converging. Furthermore, there are occasional sharp discontinuities, which is expected from a distribution of lengths with thick tails that are associated with the infinite conditional mean.[27] Meanwhile, the tF conditional averages for the four replications are essentially on top of one another and converge relatively quickly to the conditional mean of approximately 3.55.

II.D The tF procedure: Impact on Applications

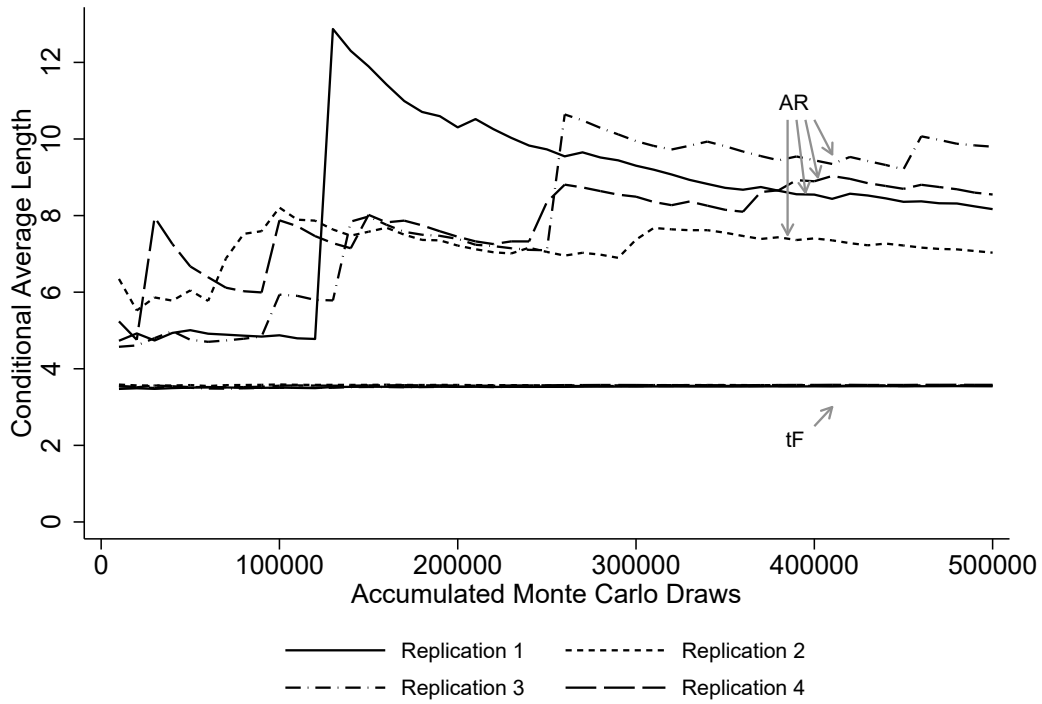
We now turn to gauging how the tF adjustments to the standard errors would impact practice, using our sample of recent *AER* papers as a guide. We take the computed or reported F-statistics from the specifications in Figure 1, and assign the corresponding adjustment factor $\frac{\sqrt{c_\alpha(F)}}{\sqrt{q_{1-\alpha}}}$. Figure 5a is the (weighted) histogram for the reciprocal of the 0.05 tF adjustment factor, which represents the degree to which the reported standard errors are understated.[28] It shows significant mass at values close to 1 (no understatement); the median reciprocal is 0.902 (understated by about 10 percent) while the 25th percentile reciprocal is 0.672 (understated by about 33 percent). The weighted mean value is 0.801, implying that the typical study is understated by about 20 percent.

Turning to the question of the magnitude of the implied inflation factors, our

[27] Recall that the Strong Law of Large Numbers states that the sample average converges to the expected value with probability one if it is finite. Furthermore, an application of the second Borel-Cantelli lemma also shows that the sample average does not converge with probability one if the population expectation is not finite.

[28] We focus on the reciprocal because the adjustment factor itself has some very large numbers. For any given study, we know that its true average will be infinite because there will always be some positive probability that $F < q_{1-\alpha}$.

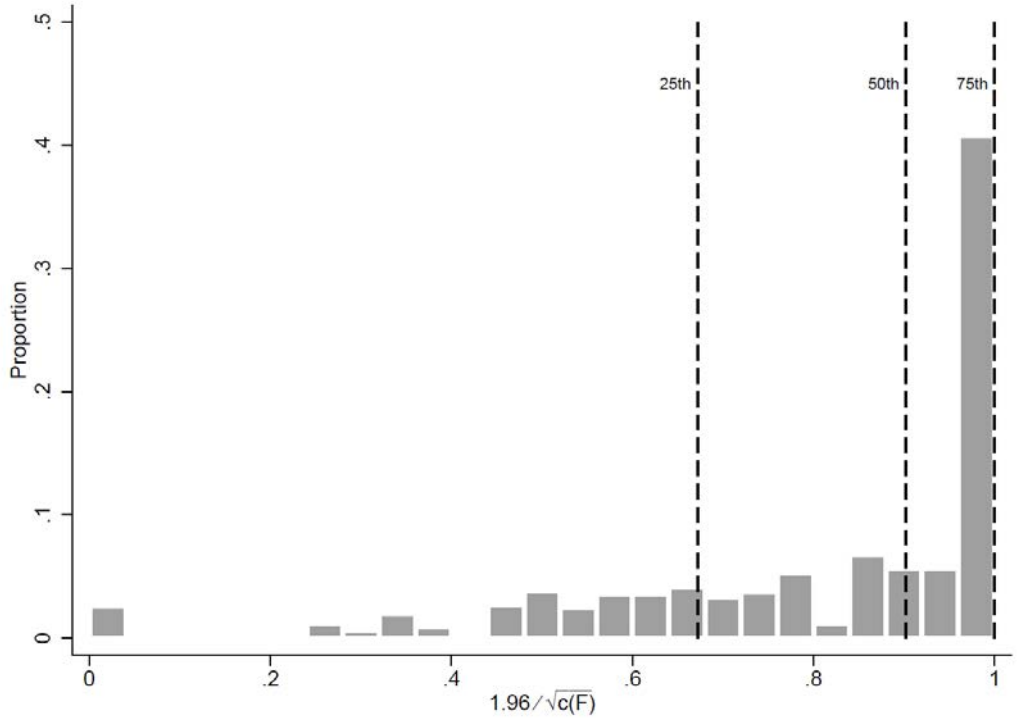
Figure 4: Monte Carlo Simulated Expected Length of tF and AR intervals, Conditional on F>1.96^2, $\rho = 0.5$, $f_0 = 3$



Note: Points on each curve represent the conditional expected length, using the specified number of accumulated Monte Carlo draws, for tF (lower four lines) and AR (upper four lines). Each of the four lines corresponds to an independent set of Monte Carlo draws.

sample of studies suggests that for one-quarter of specifications, the tF adjustment would increase confidence intervals, at a minimum, by a factor of $1/0.672 \approx 1.49$, i.e., tF confidence intervals would be at least about 50 percent wider. To understand this magnitude, it is helpful to recall that conventional 99 percent confidence intervals are about 57 percent longer than 90 percent confidence intervals. Another basis of comparison comes from our examination of a small subset of the studies for which we could obtain the microdata. For those studies that used clustered standard errors, we computed non-clustered standard errors and found that the clustered standard errors were about 25 percent larger. We conclude from these comparisons that, in practice, ignoring the tF adjustment would be an error roughly equivalent to using a 90 percent confidence interval while calling it a 99 percent confidence

Figure 5a: Distribution of $\frac{1.96}{\sqrt{c_{0.05}(F)}}$ for AER sample



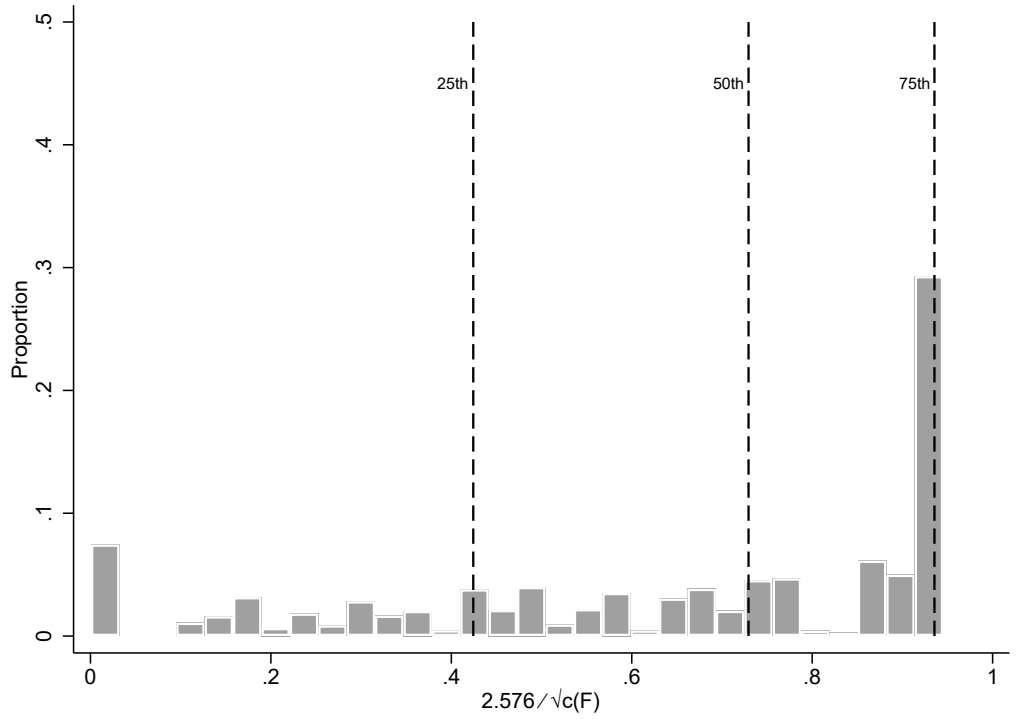
$N = 847$ specifications. The x-axis is the ratio of $\Phi^{-1}(0.975)$ to the F-dependent value $\sqrt{c_{0.05}(F)}$. All specifications use the derived F statistic and when not possible, the reported F statistic from the paper. The 6 specifications that report (rounded) first-stage standard errors of zero are excluded. Proportions are weighted; see notes to Table 1. Dashed lines correspond to the (weighted) 25$^{\text{th}}$ (0.672), 50$^{\text{th}}$ (0.902), and 75$^{\text{th}}$ (1.00) percentiles of the distribution.

interval, or substantially more severe than using non-clustered standard errors when clustered standard errors are appropriate.

Figure 5b repeats the exercise for the 0.01 tF adjustments and finds more significant degrees of adjustment: in one-quarter of the specifications, the tF adjustment would be expected to increase confidence intervals by at least a factor of 2.36, and the median adjustment factor would be 1.38.

Finally, to gauge how assessments of statistical significance are likely to be impacted by the use of the tF critical value function, Figure 6 plots all of the specifications from Table 2 in t^2, F space (using the one-to-one transformations $\frac{t^2/1.96^2}{1+t^2/1.96^2}$

Figure 5b: Distribution of $\frac{2.576}{\sqrt{c_{0.01}(F)}}$ for AER sample



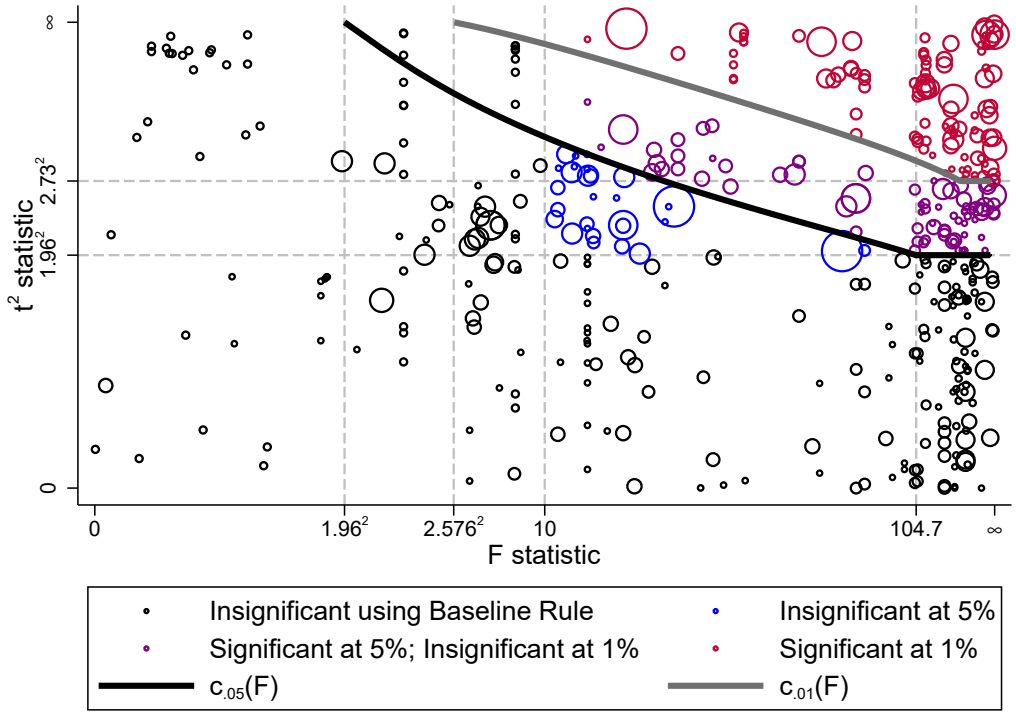
$N = 847$ specifications. The x-axis is the ratio of 2.576 to the F-dependent value $\sqrt{c_{0.01}(F)}$. See notes to Figure 5a. Dashed lines correspond to the (weighted) 25$^{\text{th}}$ (0.424), 50$^{\text{th}}$ (0.727), and 75$^{\text{th}}$ (0.936) percentiles of the distribution.

and $\frac{F/10}{1+F/10}$ for the vertical and horizontal scales to allow visualization of the full range of those statistics). It also plots the tF critical value functions for the 5 percent (black) and 1 percent (gray) levels of significance.[29] The size of each circle is proportional to the share of total specifications from the same study. The black dots represent the specifications that have a relatively low F-statistic (<10) or that have t^2 less than 1.96^2. Arguably, under current practice, researchers would have

[29]For this exercise, we further restricted the sample of specifications to those where the reported sample size for the first-stage was identical to the reported sample size for the 2SLS estimate. We have observed that it is quite common for researchers to report first-stage regressions and F statistics on samples that do not match (typically they are larger) the samples used for the 2SLS regression. The graph and the numbers reported below are quite similar if we do not make this additional restriction.

generally viewed the black circles as statistically insignificant estimates by virtue of either the observed t-ratio or the F-statistic.[30] While most of these black circles would remain insignificant using the tF adjustment, at the 5% level, some, by being above the tF critical value function would become significant.

Figure 6: Statistical Significance in AER sample, using $c_{0.05}(F)$ and $c_{0.01}(F)$



$N = 439$ specifications. Vertical scale is $\frac{t^2/1.96^2}{1+t^2/1.96^2}$ and horizontal scale is $\frac{F/10}{1+F/10}$. Size of each circle is proportional to the weight described in Table 1. The solid black and gray lines are critical value functions $c_{0.05}(F)$ and $c_{0.01}(F)$, respectively. The black circles denote cases where $t^2 < 1.96^2$ or $F < 10$. The blue circles represent those that are not significant using $c_{0.05}(F)$. The purple circles represent those that are significant at the 5 percent level using $c_{0.05}(F)$ but are not significant at the 1 percent level using $c_{0.01}(F)$. The red circles represent those that are significant at the 1 percent level using $c_{0.01}(F)$.

The remaining specifications (blue, purple, and red circles), under current norms, would most likely have been viewed as statistically significant. Of these, 24 percent

[30]We use the threshold 10 here not because it is a special threshold with respect to the theory regarding size distortions. Instead, we use it because 10 appears to be the most commonly referenced threshold in applied work.

(the blue circles) are in fact statistically insignificant at the 5 percent level, when the tF critical values are applied; the remaining 76 percent (purple and red circles) remain significant at the 5 percent level.

The proportional impact of the adjustments is larger for a higher standard for statistical significance, the 1 percent level. That is, among the specifications such that $\hat{t}^2 > 2.73^2, \hat{F} > 10$—which arguably would have commonly been interpreted as statistically significant at the 1 percent level—about 34 percent of them are statistically insignificant after applying the tF critical value function.

Although it is beyond the scope of our paper to suggest whether any of the overall conclusions of the studies in our sample would be altered in light of these adjustments, we do conclude that the tF adjustments could be expected to make a nontrivial difference in inferences made in applied research—in some cases not making much of a difference at all, but in other cases making a large difference.

Finally, we note that if the only hypothesis of interest is the null that the coefficient of interest is equal to zero, then one can simply conduct a test of whether the reduced form coefficient (in the regression of Y on Z) is zero. Indeed, this is equivalent to the AR test. On the other hand, if there is an interest in computing confidence intervals, then one requires information contained in the first-stage regression (which is used by both AR and tF).

II.E *A Priori* Restrictions on ρ

The conventional frequentist approach to statistical inference requires, by definition, that for a test at the 5 percent level of significance, the maximum rejection probability under the null hypothesis over all possible values of nuisance parameters is 0.05. We follow this conventional approach and ensure that the tF procedure is valid for any possible value of $E[F]$ and ρ.[31] While the particular values $|\rho| = 1$ are useful in derivations to provide a worst case, valid inference applies to all values of ρ between -1 and 1. Thus, for the just-identified IV model, being agnostic

[31] Our setting allows for heteroskedastic, clustered, and/or autocorrelated errors. Nevertheless, the parameter ρ simplifies to the usual endogeneity coefficient $Corr(Y - X\beta, X - \pi Z)$ which practitioners have in mind if errors are (conditionally on Z) independent and homoskedastic.

about $E[F]$ and ρ is a requirement for practitioners who wish to rely solely on the textbook *IV* assumptions that $C(Z,u) = 0$ and $C(Z,X) \neq 0$.

Adding restrictions beyond the textbook IV assumptions, for example, with *a priori* information on the parameter ρ, is possible. As referenced in Subsection III.A, one could ask, "What additional assumption about ρ could be imposed on the data generating process to allow the ± 1.96 critical values to deliver a valid 5-percent test?"

Both Lee et al. (2020) and Angrist and Kolesár (2021) calculate that using 1.96 critical values delivers a valid 5 percent test as long as one additionally assumes that ρ is less than 0.565 in absolute value.[32] A researcher's choice between adopting the conventional frequentist approach (i.e., adjusting the standard errors via tF, or via *AR* inference) or *a priori* assuming that $|\rho| \leq 0.565$ (i.e. leaving the 2SLS standard errors unadjusted) ultimately does not follow from any econometric result; instead it rests entirely on how comfortable one is with those additional *a priori* assumptions.

The plausibility of any restriction on ρ depends on the specific context. Angrist and Kolesár (2021) provide three examples in which they argue for making the $|\rho| < 0.565$ assumption. Using bounds on $|\rho|$ larger than 0.565 is also possible, which changes the interpretation: as Angrist and Kolesár (2021) point out, the ± 1.96 critical values, and assuming that $|\rho| \leq 0.76$ corresponds to a 10-percent level of significance.[33] In Appendix A.8.1, we provide the necessary inflation factors to the ± 1.96 critical values to achieve 5-percent and 1-percent levels of significance for bounds like 0.76 and other $|\rho|$ bounds between 0.565 and 1.

A separate and open empirical question is what magnitudes of ρ one might expect to see in practice. It is of course impossible to make definitive quantitative statements about the true magnitude of ρ or β, since they are both unknown parameters; also a full meta-analysis is beyond the scope of this paper. Nevertheless, as discussed in Appendix A.8.3, it is possible to use data to obtain a valid *confidence set* on ρ. The data from our sample of *AER* studies show that 1) the confidence

[32]Note that the necessary bound on $|\rho|$ depends on the desired signifiance/confidence level. For example, if 1/99 percent signifiance/confidence is intended using the nominal 2.576 critical value for the t-ratio, then the necessary bound on $|\rho|$ is 0.435, as reported in Lee et al. (2020).

[33]A 5-percent rejection rate with a precisely quantified over-rejection distortion of 5 percent means, by definition, a 10 percent test.

intervals for ρ include a broad range of values, with 24 percent of the specifications including values as large as 0.9 in absolute value, and that 2) in 18 percent of the specifications, the data would have rejected the hypothesis that $|\rho| \leq 0.565$. Appendix A.8.2 also points out that assuming $|\rho| \leq 0.565$ is equivalent to placing bounds on β.[34] For this same sample, about 30 percent of the time, assuming $|\rho| < 0.565$ is tantamount to assuming *a priori* that β is not equal to zero.

Whether or not one explores specific restrictions on ρ, it seems both costless—and not overly cautious—to report the tF standard error or confidence intervals (or AR confidence sets) as a standard inference benchmark. Such a benchmark is aligned with relying solely on the traditional IV assumptions, and also allows one to assess the gains in precision that come from imposing an assumption like $|\rho| \leq 0.565$.

III Derivation of Theoretical Results

This section explains how we derive all of the theoretical results discussed in Section II. Subsection III.A introduces the notation and shows how to analytically compute the rejection probabilities for rules that use the t-ratio, whether it be for rules like $t^2 > q_{1-\alpha}$, or $t^2 > c^*, F > F^*$, or $t^2 > c_\alpha(F)$. We do this for the case when the null hypothesis is true (for analyzing size control) and for when the alternative is true (for analyzing power). Subsection III.B defines the tF critical value function, formally states some of its properties, and describes relevant proofs. Subsection III.C formally states the results on the conditional expected length of the AR and tF confidence sets and describes relevant proofs. The details of all of the proofs of the results of this Section can be found in the Online Appendix.

[34] As noted by Van de Sijpe and Windmeijer (2021) this follows from the definitions of the reduced form and structural covariance matrices. See their equation (7) and the discussion in their Section 4.

III.A Notation and Preliminaries: Rejection probabilities for t-ratio-based rules

We begin by introducing some additional notation. Define

$$\hat{t}_{AR}(\beta_0) \equiv \frac{\hat{\pi}\left(\hat{\beta}-\beta_0\right)}{\hat{\text{se}}\left(\hat{\pi}\left(\hat{\beta}-\beta_0\right)\right)} = \frac{\hat{\pi}\left(\hat{\beta}-\beta_0\right)}{\sqrt{\hat{V}_N\left(\widehat{\pi\beta}\right) - 2\beta_0\hat{C}_N\left(\widehat{\pi\beta},\hat{\pi}\right) + \beta_0^2\hat{V}_N(\hat{\pi})}}$$

$$\hat{u}_0 = (Y - X\beta_0) - Z\hat{\pi}\left(\hat{\beta}-\beta_0\right)$$

$$\hat{\rho}(\beta_0) \equiv \frac{\hat{C}(Z\hat{u}_0, Z\hat{v})}{\sqrt{\hat{V}(Z\hat{u}_0)}\sqrt{\hat{V}(Z\hat{v})}}$$

where β_0 is a hypothesized value for β and $\hat{t}_{AR}(\beta_0)$ is a "t-ratio form" of the statistic of Anderson and Rubin (1949), so that $\hat{t}_{AR}^2(\beta_0) = AR$. $\hat{V}_N\left(\widehat{\pi\beta}\right), \hat{C}_N\left(\widehat{\pi\beta},\hat{\pi}\right)$, and $\hat{V}_N(\hat{\pi})$ are elements of the estimator for the variance-covariance matrix of the reduced form and first-stage estimators $\widehat{\pi\beta}$ and $\hat{\pi}$, respectively. $\hat{u}_0$ is the "AR residual", i.e., the residual from regressing $Y - X\beta_0$ on Z. Turning to the notation for $\hat{\rho}(\beta_0)$, note first that as we explain further in Appendix A.1, $\hat{V}(\cdot)$ and $\hat{C}(\cdot)$ (i.e., without a subscript of N) denote estimators of the middle or "meat" part of "sandwich"-type variance estimators. This allows our approach to flexibly accommodate various HAC approaches such as heteroskedastic or autocorrelated errors or one- or two-way clustering. As examples of this notation, if we consider the homoskedastic case, $\widehat{\rho}(\beta_0)$ is just the empirical correlation between the AR residual and the first-stage residual; in the heteroskedastic case, it is the same but after multiplying both residuals by the instrument.

A key equation in our analysis is

$$\hat{t}^2 = \frac{\hat{t}_{AR}^2(\beta_0)}{1 - 2\hat{\rho}(\beta_0)\frac{\hat{t}_{AR}(\beta_0)}{\hat{f}} + \frac{\hat{t}_{AR}^2(\beta_0)}{\hat{f}^2}}$$

which is a numerical equivalence that can be shown using the definitions above and with some re-arrangement of terms, as shown in Appendix A.4.

From these definitions and the above relationship, it is shown that under the weak-IV asymptotics of Staiger and Stock (1997), we obtain

$$\hat{t}^2 \xrightarrow{d} t^2 = t^2\left(t_{AR}\left(\beta_0\right), f, \rho\left(\beta_0\right)\right) \equiv \frac{t_{AR}^2\left(\beta_0\right)}{1 - 2\rho\left(\beta_0\right)\frac{t_{AR}(\beta_0)}{f} + \frac{t_{AR}^2(\beta_0)}{f^2}}, \tag{2}$$

where

$$\begin{pmatrix} t_{AR}\left(\beta_0\right) \\ f \end{pmatrix} \sim N\left(\begin{pmatrix} f_0 \frac{\Delta(\beta_0)}{\sqrt{1+2\rho\Delta(\beta_0)+\Delta^2(\beta_0)}} \\ f_0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\left(\beta_0\right) \\ \rho\left(\beta_0\right) & 1 \end{pmatrix}\right) \tag{3}$$

$$\Delta\left(\beta_0\right) = \frac{\sqrt{V\left(Zv\right)}}{\sqrt{V\left(Zu\right)}}\left(\beta - \beta_0\right) \text{ and } \rho\left(\beta_0\right) = \frac{\rho + \Delta\left(\beta_0\right)}{\sqrt{1 + 2\rho\Delta\left(\beta_0\right) + \Delta^2\left(\beta_0\right)}},$$

where $\rho = C(Zu, Zv)/\sqrt{V(Zu)V(Zv)}$ is the population correlation between Zu and Zv.[35] Thus, the squared t-ratio will converge in distribution to a random variable t^2, which is itself a function of the random variables $t_{AR}\left(\beta_0\right)$ and f, which are themselves jointly bivariate normal with unit variances and correlation $\rho\left(\beta_0\right)$. Note that when the null hypothesis is true, $\beta = \beta_0$ implies that $\Delta\left(\beta_0\right) = 0$ and $\rho\left(\beta_0\right) = \rho$.

These relationships hold true for error structures that depart from i.i.d., but when we consider the specific case of homoskedasticity, the formula in (2) can be shown to yield equation 2.22 in Stock and Yogo (2005).

Remark. The econometric literature has long established the existence of distortions in inference that occur when using the t-ratio for IV. Equation (2) is yet another way to see the same result. Specifically, the conventional asymptotic approximation implicitly treats t^2 as a chi-squared with one degree of freedom—which is the distribution of the numerator in (2) and therefore, essentially, ignores the denominator in (2) by treating f as infinite. But, as Figure 1 illustrates, in our sample of studies, half of the time $\hat{F} = \hat{f}^2$ is less than 46.

We use the expressions above to compute rejection probabilities for different test procedures that reject the null hypothesis when $t^2 > k(F)$, where $k(F)$ is a

[35] In the display, to simplify the presentation, we present notation for $\Delta(\beta_0)$ for the heteroskedastic case rather than the most general HAC case. Details of these derivations extended to the general HAC case are contained in the Online Appendix.

general critical value function that could depend on F:

$$\begin{aligned}
\text{Conventional } t\text{-ratio test: } & k(F) = q_{1-\alpha} \\
\text{Single } F \text{ threshold test: } & k(F) = \begin{cases} c^* & \text{if } F > F^* \\ \infty & \text{if } F \leq F^* \end{cases} \\
tF \text{ critical value function: } & k(F) = c_\alpha(F)
\end{aligned}$$

In all cases, the rejection probability can be expressed as

(4)

$$\begin{aligned}
\Pr_{\Delta(\beta_0),\rho,f_0}\left[t^2 > k(F)\right] = \int\int 1\left[t^2(x,y,\rho(\beta_0)) > k\left(y^2\right)\right] \\
\times \varphi\left(x - f_0\frac{\Delta(\beta_0)}{\sqrt{1+2\rho\Delta(\beta_0)+\Delta^2(\beta_0)}}, y - f_0; \rho(\beta_0)\right) dxdy
\end{aligned}$$

where $1[\cdot]$ is the indicator variable, and $\varphi(\cdot,\cdot;r)$ is the bivariate normal density with means zero, unit variances, and correlation r.

This expression allows us to compute rejection probabilities up to the accuracy of numerical integration. We use these computations to 1) illustrate the magnitude of inferential distortions caused by the usual t-ratio procedure (Figure 2 Panel (A)), 2) verify that the tF critical value function controls the significance level, as intended (Figure 2 Panel (B)), and 3) construct power functions (Figure 3 and Appendix Figure A2).[36]

Remark. In addition, expression (4) also allows us to answer the following questions: 1) What restrictions on the nuisance parameter space f_0, ρ could one impose so that the usual t-ratio procedure has the intended significance level?[37] 2) For single threshold rules, what minimal threshold for F^* could one use if c^* is set to the nominal value $q_{1-\alpha}$? and 3) How do these answers change for different significance levels? Appendix A.7 (and a previous version of our paper, Lee et al.

[36]Note that it is straightforward to use the mean shift in $t_{AR}(\beta_0)$ from expression (3) to compute the power function for AR.

[37]Kocherlakota (2020) develops a method that incorporates nuisance parameter information in a t-ratio test.

(2020)) provides answers to these questions.

III.B Construction of the tF critical value function

Our objective is to obtain a critical value function $c_\alpha(F)$ that smoothly adjusts according to the first-stage F-statistic and that also controls size, i.e., it has the property that

$$\Pr_{\Delta(\beta_0)=0,\rho,f_0}\left[t^2 > c_\alpha(F)\right] \leq \alpha$$

for all ρ and $f_0 \neq 0$. Deferring details to Appendix B, we now outline the construction of such a function $c_\alpha(F)$, which – as is apparent from Tables 3 Panel A and 3 Panel B – consists of an initial strictly decreasing segment ranging from $q_{1-\alpha}$ to some point, followed by a flat function beyond that point. This plateau structure is motivated by practical considerations, since researchers may desire a constant critical value function as long as the F statistic is large enough.

The first step of our construction – the decreasing segment of the critical value function – stems from the conjecture of Stock and Yogo (2005) that for small, fixed values of f_0 (when instruments are "weak"), the "worst case" null rejection probability occurs when $\rho = \pm 1$.[38] This leads to obtaining a function $c_\alpha(F)$ such that the null rejection probability under $\rho = \pm 1$ is exactly equal to α,

$$\Pr_{\Delta(\beta_0)=0,|\rho|=1,f_0}\left[t^2 > c_\alpha(F)\right] = \alpha \tag{5}$$

for some set of small values of f_0. To simplify exposition, we focus on the case where $\rho = 1$ and f_0 positive.[39]

The following fact is central to our construction of the tF critical value function: when $\rho = 1$, t_{AR} is a linear function of f, and therefore Equation (2) reduces to

$$t^2 = \frac{f^2(f - f_0)^2}{f_0^2}$$

[38] In Appendix B, we substantiate this conjecture.

[39] However, all the discussion below for $\rho = 1$ and f_0 positive applies symmetrically for $\rho = -1$ and/or f_0 negative.

which is a quartic function in f, uniquely indexed by the single parameter f_0. This quartic function has the shape of a "W", with one trough located at $f = 0$, the other trough at $f = f_0$, and an interior peak at $f = f_0/2$. Furthermore, the magnitude of the location and height of the interior peak of the "W" function is monotonically increasing in $|f_0|$. Three examples of this "W"-shaped quartic function are illustrated in Figure 7, which plots t^2 as a function of f as the blue, red, and gray curves, corresponding to three values of f_0, labeled f_0', f_0'', and f_0'''.[40]

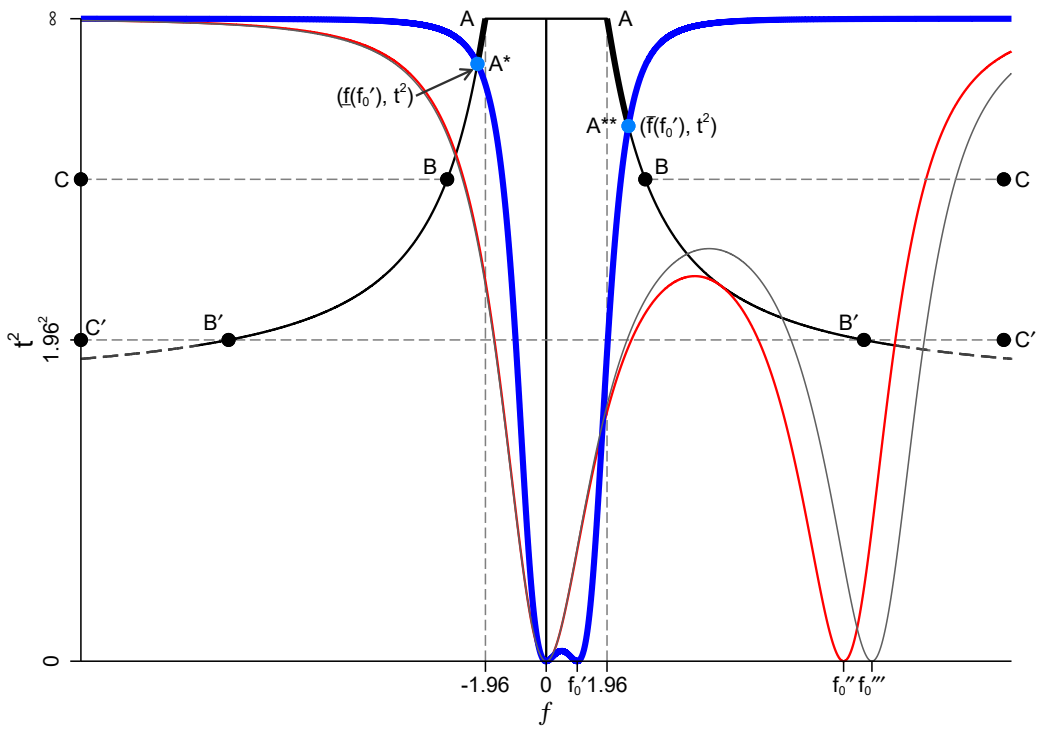
The case of $\rho = 1$ greatly simplifies the expression of the null rejection probability for any critical value function (as in Equation (4)). The expression now involves a single random variable, f, which is normally distributed with mean f_0 and unit variance. That is, we can now characterize the null rejection probability by the probability that f takes on a value for which the quartic t^2 curve is above the critical value function. For any continuous and decreasing (in f^2) critical value function (that eventually plateaus), there exists an interval of values of f_0 for which the "W" curve and the critical value function intersect only twice. The acceptance probability is then simply $\Phi\left(\bar{f}\left(f_0\right) - f_0\right) - \Phi\left(\underline{f}\left(f_0\right) - f_0\right)$, where the intersections between the two curves are denoted by $\bar{f}\left(f_0\right)$ and $\underline{f}\left(f_0\right)$. For example, for the blue curve in Figure 7 (corresponding to $f_0 = f_0'$) the acceptance probability is equal to the probability that f lies in the interval given by $[\underline{f}\left(f_0'\right), \bar{f}\left(f_0'\right)]$.[41]

We use this simple form of the acceptance region to define a decreasing function $\tilde{c}_\alpha(\cdot)$, which will be coincident with the eventual critical value function. Specifically, we seek a decreasing function (in f^2) that intersects each of the "W" functions (indexed by f_0) at two points, $\bar{f}\left(f_0\right)$ and $\underline{f}\left(f_0\right)$,where $\Phi\left(\bar{f}(f_0\right) - f_0) - \Phi(\underline{f}\left(f_0\right) - f_0) = 1 - \alpha$. This definition can be expressed more formally as the function $\tilde{c}_\alpha(\cdot)$

[40]The figure uses the transformation $\frac{(t^2/1.96^2)}{1+(t^2/1.96^2)}$ for the vertical axis to aid visualization of the curves.

[41]Figure 7 also shows that, for large values of f_0, the rejection region is not necessarily an interval, such as for the gray curve (represented by $f_0 = f_0'''$).

Figure 7: Construction of the tF Critical Value Function



Note: To aid visualization of values, vertical axis uses the transformation $\frac{(t^2/1.96^2)}{1+(t^2/1.96^2)}$.

satisfying the following system of equations:

$$
\begin{aligned}
\frac{\bar{f}(f_0)^2\left(\bar{f}(f_0)-f_0\right)^2}{f_0^2}-\tilde{c}_\alpha\left(\bar{f}(f_0)^2\right)&=0 \\
\Phi\left(\bar{f}(f_0)-f_0\right)-\Phi\left(\underline{f}(f_0)-f_0\right)&=1-\alpha \\
\frac{\underline{f}(f_0)^2\left(\underline{f}(f_0)-f_0\right)^2}{f_0^2}-\tilde{c}_\alpha\left(\underline{f}(f_0)^2\right)&=0
\end{aligned}
\tag{6}
$$

for a set of small values of f_0.

Whether or not there exists any continuous and decreasing function $\tilde{c}_\alpha\left(f^2\right)$ satisfying this system of equations is not obvious and is technically challenging to prove. We defer those details to Appendix B. Here, we apply the results in the appendix to illustrate how we construct the desired critical value function.

We are able to obtain a "local" solution to (6) as the critical value function increases without bound, which occurs as $f^2 \downarrow q_{1-\alpha}$. In particular, from Lemma 9(i) in Appendix B, as $f^2 \downarrow q_{1-\alpha}$, the function $\tilde{c}_\alpha(f^2)$ behaves as

$$\tilde{c}_\alpha(f^2) = \frac{q_{1-\alpha}^3}{f^2 - q_{1-\alpha}} - \left(3q_{1-\alpha} - \frac{q_{1-\alpha}^2}{2} + \frac{q_{1-\alpha}^3}{6}\right) + O\left(\sqrt{f^2 - q_{1-\alpha}}\right). \tag{7}$$

This equation is derived from applying a theorem from Fefferman (2021).[42]

With equation (7) in hand, constructing the decreasing part of the tF critical value function is straightforward. We provide a graphical explanation of the procedure in Figure 7, focusing on the leading case of $\alpha = 0.05$.[43] We start with a set of points $(\underline{f}, \tilde{c}_{0.05}(\underline{f}^2))$, defined over the small interval $\underline{f} \in [-1.96 - \varepsilon, -1.96)$ for $\varepsilon > 0$. This interval is motivated by the theoretical result in equation (7), and specifically is based on that equation's leading terms. For each point over that interval, the third equation in (6) can be used to solve for f_0, allowing $\underline{f}$ to be relabeled $\underline{f}(f_0)$. Then, the second equation in (6) can be used to solve for $\bar{f}(f_0)$. Finally, the first equation in (6) can be used to solve for $\tilde{c}_{0.05}(\bar{f}(f_0)^2)$. This mapping produces a segment of the function defined on an interval that is longer than $[-1.96 - \varepsilon, -1.96)$. Due to symmetry of the function $\tilde{c}_{0.05}(f^2)$ in f, one can use this extended version of the function, $\tilde{c}_{0.05}(f^2)$, as a new starting segment and repeat the process.[44] Figure 7 illustrates one iteration of this mapping starting with the shorter critical value function segment given by the segment AA^* on the left and the corresponding extended segment given by the segment AA^{**} on the right. These segments terminate at the blue "W" function and show how the blue endpoint on the left (A^*) maps to the blue endpoint on the right (A^{**}).

This process can be iterated to produce incremental extensions to the curve $\tilde{c}_{0.05}(\cdot)$, as long as the associated "W" curves intersect the extended curves only

[42] See also Baldomá et al. (2007) and Baldomá, Fontich and Martín (2020).

[43] The "W" functions in Figure (7) do not depend on α, but the critical value curves do. The approach we outline here for $\alpha = 0.05$ can be applied more generally for other values of α. In Appendix B, we give details regarding other values of α and discuss some differences that may arise.

[44] With some abuse of notation, we will use $\tilde{c}_\alpha(\cdot)$ to refer to both the original function that exists according to Lemma 9 as well as every extension of that function as described above.

twice. It traces out a decreasing segment until the curve terminates at a very specific endpoint – where the "W" curve, whose "right arm" passes through the endpoint, posseses an interior "hump" that is tangent to the critical value function, as depicted in Figure 7 by the red curve, corresponding to $f_0 = f_0''$. Therefore, the set of values $|f_0| \leq f_0''$ are precisely the "set of small values" of f_0 (referenced above) for which Equation (5) and the system of equations (6) hold. In principle, one could alternatively attempt to extend the decreasing segment further (as illustrated by the extended dashed line in the figure), so that the rejection probability continued to be equal to 0.05 for $f_0 > f_0''$. It is clear, however, that for such values (e.g., $f_0 = f_0'''$ in the figure) the associated "W" curve (e.g. the gray curve) will intersect the critical value function more than two times, and therefore the system (6), which presumed two intersection points, could not be used. We do not attempt this extension for technical reasons explained in greater detail in Appendix B.1.

The second, and more straightforward, step in constructing the tF critical value function is to determine where, along function $\tilde{c}_\alpha\left(f^2\right)$, the critical value function plateaus. There are many candidates: for example, one could use the piece-wise function that passes through ABC in the figure, where the segment passing through BC could potentially start on any point on the (thick or thin) solid black line. Among all the plateaus that control size, the choice of a lower plateau will lead to a more powerful test; therefore, we define the tF critical value function $c_\alpha(F)$ to be the one with the lowest possible plateau that controls rejection probabilities to be less than or equal to α, for all values of ρ and f_0. In practice, we use numerical integration of the expression in (4) to compute these rejection probabilities, as illustrated in Figure 2, for a grid of values for f_0 and ρ to verify size control.[45] For $\alpha = 0.05$, our numerical analysis indicates that size is controlled to be 0.05 when the critical value function, represented by the function that passes through $AB'C'$ in Figure 7, has the plateau level set to equal the chi-square critical value 1.96^2. For $\alpha = 0.01$, the construction of the decreasing segment of the critical value ends before the function falls below the chi-square critical value of 2.58^2. For that case, the plateau is set

[45]Specifically, we use two grids. The first grid consists of values of ρ that range from 0 to 1 in increments of 0.01, and values of f_0 that range from 0 to 80 in increments of 0.25. The second grid is one that focuses on the ρ values of 0.995, 0.996, 0.997, 0.998, and 0.999 and f_0 values that range from 0 to 80 at increments of 0.01.

to the smallest possible value of that construction process, 2.73^2. Appendix B.3 provides a step-by-step algorithm for obtaining the entire tF critical value function as outlined above.

The construction of the critical value function implies the existence of an entire class of critical value functions that also control size, with decreasing segments that are all coincident, with the only difference being where the plateau begins; the tF critical value function, by definition, has the lowest plateau. A natural question to ask is whether, for a given plateau, there might exist alternative critical value functions that also control size, with a similar structure, but distinct from the decreasing segment that we have constructed. In Appendix B, we specify a set of properties that critical value functions could possess, and show that the class of critical value functions described above is the only class that satisfies those properties.

III.C Conditional Expected Length: AR and tF confidence sets

This subsection describes how we obtain our results on the conditional expected length of AR and tF intervals. Our motivation to examine expected length stemmed from the traditional power curve analysis in Subsection II.C, which showed that neither AR nor tF seemed to dominate across all values of $\Delta(\beta_0)$ or differing combinations of ρ and f_0. A natural summary measure of power is that of expected length of the confidence set, which has the equivalent interpretation, due to Pratt (1961), as the average Type II error, where the averaging occurs across all possible false hypotheses β_0, where each value of β_0 in the parameter space is given equal weight. Power curves are conceived as rejection rates while keeping β_0 fixed while varying β, but our curves, since they are functions of $\Delta(\beta_0) = \frac{\sqrt{V(Zv)}}{\sqrt{V(Zu)}}(\beta - \beta_0)$, could equivalently be viewed as graphing power fixing β, while varying β_0. So the expected length of the confidence set is equivalent to averaging 1 minus power, averaging across $\Delta(\beta_0)$.

Examining *unconditional* expected length, however, will not be informative since we know, from Dufour (1997), that inverting both the AR and tF tests, by virtue of delivering correct confidence levels, will have infinite unconditional expected length. Thus, we turn to examining the expected length of confidence sets

conditional on $F > q_{1-\alpha}$. The event $F > q_{1-\alpha}$ is important because it is the necessary and sufficient condition for both the *AR* and tF confidence sets to be bounded intervals; they have unbounded confidence sets with identical probabilities. This allows us to interpret the conditional expected length as the average Type II error—averaged across all false hypotheses β_0—conditional on the confidence set being an interval. Furthermore, conditional expected length is likely to be of interest to practitioners who may wonder if they should expect *AR* or tF intervals to be shorter.

Given the ambiguity in the power comparison results, it was surprising to find that an expected length comparison yields a stark contrast and clearly dominant method: tF intervals are shorter in expectation. Indeed, we reach a somewhat stronger result. The conditional expected length for the *AR* confidence interval is infinite, while the conditional expected length of the tF interval is finite.

More formally, what we establish is the following. In any finite sample, there are three confidence interval lengths that are relevant to this result, namely $\hat{L}_{IV}$ (the length of the conventional t-ratio-based confidence interval), $\hat{L}_{AR}$, and $\hat{L}_{tF}$, (the lengths of the *AR* and tF intervals, respectively) and each of these converge in distribution to random variables L_{IV}, L_{AR}, and L_{tF}, respectively. Appendices C.2 and C.3 show that for all $\rho, f_0 \neq 0$

$$E\left[L_{AR}|F > q_{1-\alpha}\right] = \infty \quad \text{and} \quad E\left[L_{tF}|F > q_{1-\alpha}\right] < \infty.$$

We next provide some intuition for this result. We show in Appendix C.1 that conditional on $F > q_{1-\alpha}$, we can write L_{AR} and L_{tF} as inflated versions of L_{IV}, i.e.,

$$\begin{aligned} L_{AR} &= \frac{\sqrt{F}\sqrt{F - q_{1-\alpha}(1-\tilde{\rho}^2)}}{F - q_{1-\alpha}} L_{IV} \\ \text{and} \quad L_{tF} &= \sqrt{\frac{c_\alpha(F)}{q_{1-\alpha}}} L_{IV}, \end{aligned} \tag{8}$$

where $\tilde{\rho}^2 = \frac{(-t_{AR}(\beta)+\rho f)^2}{\left(f^2 - 2\rho t_{AR}(\beta) f + t_{AR}^2(\beta)\right)}$.

It turns out that the L_{AR} inflation factor explodes as F approaches $q_{1-\alpha}$ from above, and even accounting for the other parts of the inflation factor, the denominator $(F - q_{1-\alpha})$ leads to an infinite conditional expected length. As for L_{tF}, the

inflation factor does not grow as quickly as F approaches $q_{1-\alpha}$ from above, and in particular grows slowly enough that conditional expected length is finite. The key to this result is our finding in Appendix B.2 that

$$\lim_{F \downarrow q_{1-\alpha}} c_\alpha(F)(F - q_{1-\alpha}) = q_{1-\alpha}^3.$$

This result allows us to show integrability of $\sqrt{c_\alpha(F)}$ because it shows that

$$\sqrt{c_\alpha(F)} = \sqrt{\frac{c_\alpha(F)(F - q_{1-\alpha})}{F - q_{1-\alpha}}} \leq \frac{M}{\sqrt{F - q_{1-\alpha}}}$$

for some bound M, and in a neighborhood of $q_{1-\alpha}$, $1/\sqrt{F - q_{1-\alpha}}$ is integrable. Appendix C provides the full proof.

In summary, these results show that the expected length of the tF confidence set is (infinitely) shorter than that of the AR confidence set when $F > q_{1-\alpha}$. At the same time, when $F < q_{1-\alpha}$, the tF confidence set always consists of the entire line. By contrast, when $F < q_{1-\alpha}$, the AR confidence set is either the entire real line or, possibly, a set that consists of all values *outside* a finite interval (see discussion in Andrews, Stock and Sun (2019), Dufour and Taamouti (2005), and Mikusheva (2010)).[46] Thus, a trade-off in length is expected: tF does better when $F > q_{1-\alpha}$, but AR does better when $F < q_{1-\alpha}$. Note that the statement that tF does not dominate AR in terms of expected length depends crucially on the presumption that researchers are prepared to properly report, in the event that $F < q_{1-\alpha}$, a non-convex and unbounded confidence set.[47] If, for example, in practice researchers effectively ignore the non-convexity and simply use the whole real line as the confidence set, then the confidence sets for tF and AR would coincide when $F < q_{1-\alpha}$. In other words, the unconditional expected difference in lengths between a "convexified" AR confidence set and the tF interval would always favor tF.

[46] When $F = q_{1-\alpha}$, the tF confidence set is the entire real line, whereas the AR confidence set can be the entire real line, a left- or right-bounded interval, or the empty set.

[47] We are unaware of an example when such a non-convex confidence set is reported other than Cruz and Moreira (2005).

IV Conclusion and Extensions

Since the work of Dufour (1997), it has been known in the econometrics community that the conventional t-ratio delivers incorrect size; the work of Staiger and Stock (1997) and Stock and Yogo (2005) provided the framework and approach for quantifying—and fixing—these distortions to inference.

Yet practitioners, while using the ± 1.96 critical values that are more commonly associated with a 5 percent test or 95 percent confidence interval, seem not to have been using those results to qualify their inferences (e.g., they typically do not explicitly state that they are assuming $E[F] > 6.88$, recognizing the test as a 10 percent significance test), nor have they been precise about the consequences of incorporating the first-stage F statistic into the inferences about β, even though the literature has provided such a method (e.g., they have not explicitly described the rule, "reject if and only if $t^2 > 1.96^2, F > 16.38$," as a test at the 15 percent level of significance). Applied work also rarely uses the AR statistic, which has been known to deliver valid inference.

This paper develops a "continuous" version of the critical value functions that result from the application of Staiger and Stock (1997) to the values in Stock and Yogo (2005). This smooth adjustment approach reduces the scope for misapplication or misinterpretation since the interpretation is straightforward: after adjustment of the standard errors, hypothesis tests and interval estimates have their intended significance or confidence levels, irrespective of the true values of the nuisance parameters—just like AR.

In our comparison between the two alternatives—AR and tF—both of which have correct size, we discover a somewhat surprising fact about the AR confidence set. Conditional on the confidence set being a bounded interval, it has infinite expected length, due to the thick upper tail of the probability distribution of lengths. By contrast, the tF confidence set has finite expected length, whenever it is a bounded interval. Therefore, in addition to the tF adjustment allowing a way to re-assess the inferences of past studies, there is a practical reason for considering its use for applied work, as an alternative to AR going forward.

There are some issues that we believe are worthy of deeper investigation. The

scope of our study was limited to the common case of the single instrument IV model, but it would be natural to expect the same kinds of issues to be at play with the over-identified model, given the critical value tables of Stock and Yogo (2005), which are appropriate for over-identified models as well. In ongoing work, we are exploring the extent to which the tF approach can be applied to over-identified models.

References

Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics*, 20: 46–63.

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock. 2006. "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression." *Econometrica*, 74: 715–752.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics*, 11: 727–753.

Angrist, Joshua, and Jorn-Steffen Pischke. 2009*a*. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Angrist, Joshua, and Jorn-Steffen Pischke. 2009*b*. "A Note on Bias in Just Identified IV with Weak Instruments." `econ.lse.ac.uk/staff/spischke/mhe/josh/solon_justid_April14.pdf`, last accessed July 31, 2021.

Angrist, Joshua, and Michal Kolesár. 2021. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." NBER Working Paper 29417.

Baldomá, I., E. Fontich, and P. Martín. 2020. "Invariant Manifolds of Parabolic Fixed Points (I). Existence and Dependence on Parameters." *Journal of Differential Equations*, 268: 5516–5573.

Baldomá, I., E. Fontich, R. De La Llave, and P. Martín. 2007. "The Parameterization Method for One-Dimensional Invariant Manifolds of Higher Dimensional Parabolic Fixed Points." *Discrete & Continuous Dynamical Systems*, 17: 835–865.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables is Weak." *Journal of American Statistical Association*, 90: 443–450.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference With Multiway Clustering." *Journal of Business Economics and Statistics*, 77: 238–249.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica*, 83: 2453–2483.

Chioda, Laura, and Michael Jansson. 2005. "Optimal Conditional Inference for Instrumental Variables Regression." UC Berkeley Working Paper.

Cruz, L. M., and M. J. Moreira. 2005. "On the Validity of Econometric Techniques With Weak Instruments: Inference on Returns to Education Using Compulsory School Attendance Laws." *Journal of Human Resources*, 40: 393–410.

Dufour, Jean-Marie. 1997. "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models." *Econometrica*, 65: 1365–1388.

Dufour, Jean-Marie, and Mohamed Taamouti. 2005. "Projection-based statistical inference in linear structural models with possibly weak instruments." *Econometrica*, 73(4): 1351–1365.

Fefferman, Charles. 2021. "Invariant Curves for Degenerate Hyperbolic Maps of the Plane." *arXiv preprint arXiv:2108.04887*.

Feir, Donna, Thomas Lemieux, and Vadim Marmer. 2016. "Weak Identification in Fuzzy Regression Discontinuity Designs." *Journal of Business & Economic Statistics*, 34: 185–196.

Gleser, L. J., and J. T. Hwang. 1987. "The Non-Existence of 100(1-α)% Confidence Sets of Finite Expected Diameter in Errors-in-Variables and Related Models." *Annals of Statistics*, 15: 1351–1362.

Hall, Alastair R., Glenn D. Rudebusch, and David W. Wilcox. 1996. "Judging Instrument Relevance in Instrumental Variables Estimation." *International Economic Review*, 37: 283–298.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62: 467–475.

Keane, Michael, and Timothy Neal. 2021. "A Practical Guide To Weak Instruments." University of New South Wales Working Paper.

Kleibergen, Frank, and Richard Paap. 2006. "Generalized Reduced Rank Tests Using the Singular Value Decomposition." *Journal of Econometrics*, 133: 97–126.

Kocherlakota, Narayana R. 2020. "Analytical Formulae for Accurately Sized t-tests in the Single Instrument Case." *Economic Letters*, 189. 109053.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48: 281–355.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter. 2020. "Valid t-ratio Inference for IV." Princeton University Working Paper.

Mikusheva, Anna. 2010. "Robust Confidence Sets in the Presence of Weak Instruments." *Journal of Econometrics*, 157: 236–247.

Moreira, Humberto, and Marcelo J. Moreira. 2019. "Optimal Two-Sided Tests for Instrumental Variables Regression with Heteroskedastic and Autocorrelated Errors." *Journal of Econometrics*, 213: 398–433.

Moreira, Marcelo J. 2002. "Tests with Correct Size in the Simultaneous Equations Model." PhD diss. UC Berkeley.

Moreira, Marcelo J. 2009. "Tests with Correct Size when Instruments Can Be Arbitrarily Weak." *Journal of Econometrics*, 152: 131–140.

Nelson, C. R., and R. Startz. 1990. "The Distribution of the Instrumental Variables Estimator and Its t-Ratio when the Instrument is a Poor One." *Journal of Business*, 63: 5125–5140.

Pratt, John W. 1961. "Length of Confidence Intervals." *Journal of the American Statistical Association*, 56: 549–567.

Rothenberg, Thomas J. 1984. "Approximating the Distributions of Econometric Estimators and Test Statistics." In *Handbook of Econometrics*Vol. 2, , ed. Z. Griliches and M. D. Intriligator, Chapter 15, 881–935. Amsterdam:Elsevier Science.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 65: 557–586.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, ed. Donald W.K. Andrews and James H. Stock, Chapter 5, 80–108. Cambridge University Press.

Stock, J. H., and J. Wright. 2000. "GMM with Weak Identification." *Econometrica*, 68: 1055–1096.

Van de Sijpe, Nicolas, and Frank Windmeijer. 2021. "On the Power of the Conditional Likelihood Ratio and Related Tests for Weak-Instrument Robust Inference." University of Oxford Working Paper.