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

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 also thank Orley Ashenfelter, Marinho Bertanha, Stéphane Bonhomme, Janet Currie, Michal Kolesár, Alex Mas, Ulrich Mueller, Zhuan Pei, Mikkel PlagborgMøller, Chris Sims, Eric Talley, Mark Watson, 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. We are also grateful to Camilla Adams, Victoria Angelova, Jared Grogan, Bailey Palmer, and Myera Rashid, and especially Sarah Frick and Katie Guyot for extraordinary research assistance. We acknowledge support from the Princeton School of Public and International Affairs. 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
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 the approach for correcting inference of Stock and Yogo (2005), we introduce the *tF* critical value function, leading to a minimized standard error adjustment factor that is a smooth function of the first-stage *F*-statistic. Applying the correction to a sample of 61 AER papers leads to a 25 percent increase in standard errors, on average. *tF* confidence intervals have shorter expected length than those of Anderson and Rubin (1949), whenever both are bounded intervals.

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
jrporter@ssc.wisc.edu

Supplementary Material for "Valid t-ratio Inference for IV" is available at
http://www.princeton.edu/~davidlee/wp/SupplementarytF.html

1 Introduction

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

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

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 there-

[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, including classes of unbiased tests. (Here we are referring to the unbiasedness of the test procedure, which requires rejection probabilities under all alternatives to be larger than that under the null, as opposed to the unbiasedness of the *IV* estimator.) 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

fore 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), for example, make clear that the distortions in inference also occur in the *single instrumental variable, just-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

[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 $\hat{\beta}$ 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.

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 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, the tF critical value function is the "smallest" one in the sense that any alternative function that is weakly below the tF function everywhere (and strictly below in the decreasing part of the function) 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, 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 the subsample of 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, on average, to an increase in confidence interval length of about 25 percent. 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 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 2 uses recent papers published in the *AER* to characterize current inferential practices for the single-instrument IV model. In Section 3, we first describe the tF procedure—the critical values, the main results on power, and its application to our sample of studies. Section 4 describes how the results stated in Section 3 are derived. Section 5 concludes.

2 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]

[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, using a subset of the studies for which it was possible to obtain the original microdata.

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.

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 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.

For each configuration, Table 1 reports the number of specifications, as well as proportions (parentheses) and weighted proportions (brackets), where the weight for each specification is the inverse of the total number of specifications reported from its 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 the 2SLS and the reduced form, without reporting the first-stage F (2.4 percent).

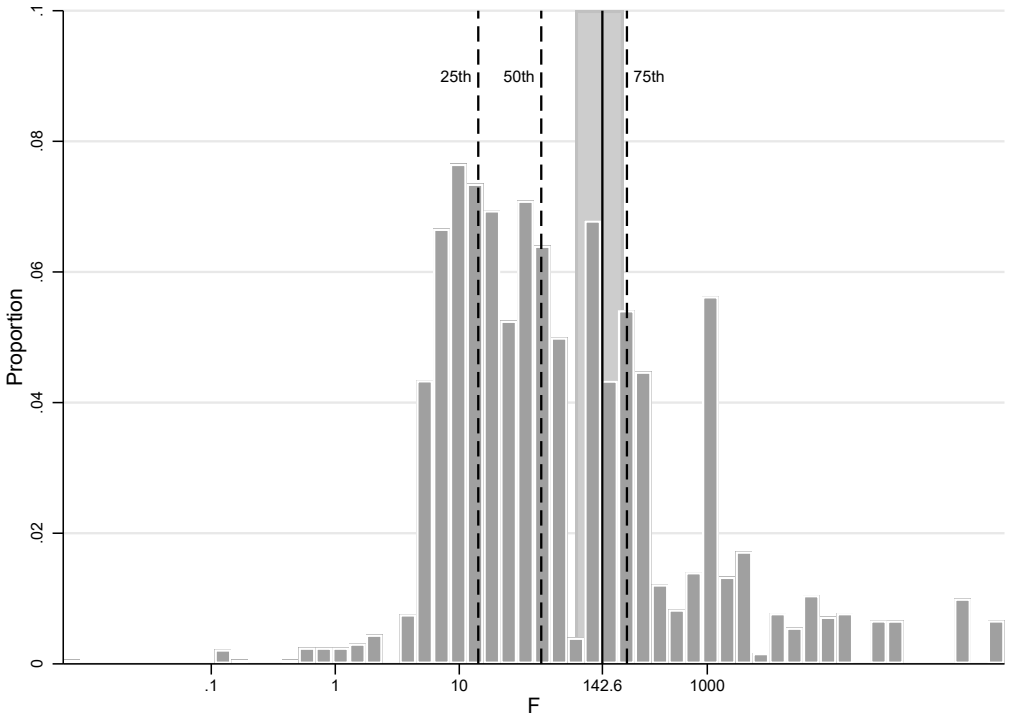
In our analysis of the data, in order to maximize the number of specifications for which we have a first-stage F-statistic, we compute it from the reported first-stage coefficients and standard errors, but whenever this is not possible, we use the explicitly reported F-statistic.[10]

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, 41.84, and 255, respectively. The figure shows that most of the reported first-stage F-statistics in these studies do pass commonly cited thresholds such as 10. 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 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

[10] 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.

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 25th (14.23), 50th (45.84), and 75th (225) percentiles of the distribution. The shaded region denotes the range between the 0.5th and 99.5th percentiles of a non-central χ^2_1 distribution with a non-centrality parameter equal to 142.6.

2SLS specification is used for a "placebo" analysis, where insignificant results are consistent with the identification strategy of the paper. 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-

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.

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.

3 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 4 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.[11] 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 application of the tF adjustments impacts the statistical inferences in our sample of *AER* studies.

[11] 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 4 that our formulas can be adapted to analyze other levels of significance or confidence levels.

3.1 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.[12]

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 $\hat{F}$.

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 un-

[12] 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.

derlies 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.[13]

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—for different values of $E[F]$ and ρ, where $E[F] = f_0^2 + 1$.[14] The figure illustrates that

[13] Strictly speaking, Stock and Yogo (2005) use a homoskedastic model in which ρ is the correlation between the errors u and v, but we use a heteroskedastic version for exposition here.

[14] As we explain in detail in Section 4, 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.

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 probabiity 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), for example, 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 $\hat{F}$. More specifically, they develop a framework for determining 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. Put equivalently, this implies a confidence interval procedure that sets the confidence interval to the entire real line if $F < F^*$, and otherwise use $\pm\sqrt{c^*} \times \text{se}$ for the confidence interval.

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

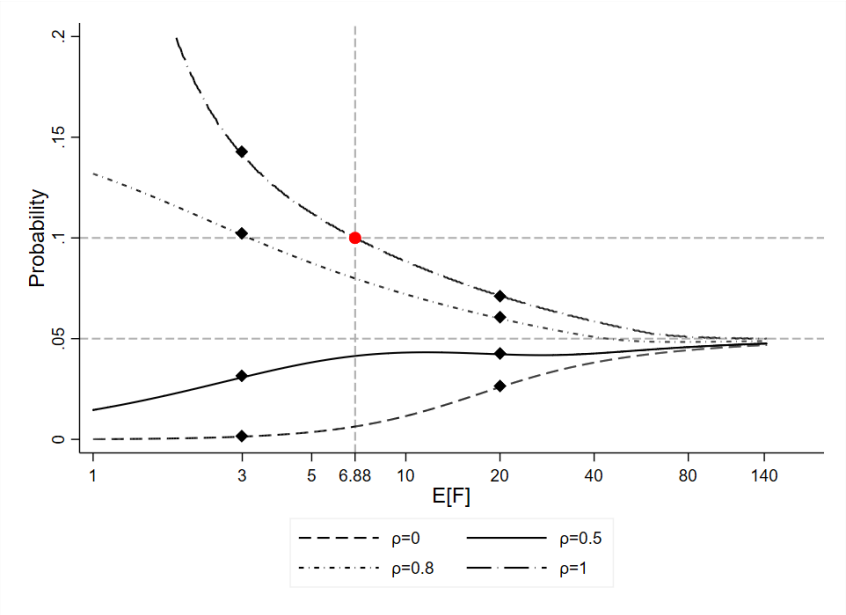
$$\Pr\left[t^2 > c_\alpha(F)\right] \leq \alpha$$

for a pre-specified significance level α, where $c_\alpha(F)$ is a smooth function of F, instead of a step function.[15] As we will show below, inference based on tF has sig-

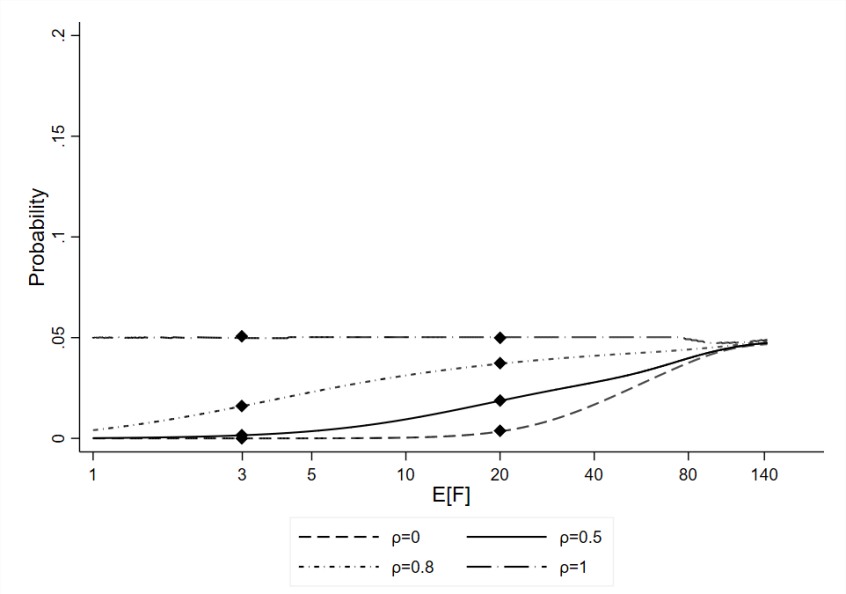
[15] As with the approach of Staiger and Stock (1997), the probability considered is an unconditional one. See Chioda and Jansson (2005) for an analysis of inference conditional on the observed

Figure 2

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



(b) Pr[$t^2 > c_{.05}(F)$] vs. E[F], for selected values of ρ



Note: Scale is ln(E[F]). Red circle 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.

nificant 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.

3.2 The tF procedure: critical values and valid inference

Table 3a reports numbers that reflect the shape of the function $c_{.05}(F)$. Specifically, each of the 100 entries is a selected value of the first-stage F statistic, along with the corresponding standard error adjustment factor, $\frac{\sqrt{c_{.05}(F)}}{1.96}$. $c_{.05}(F)$ tends to ∞ as F tends to 1.96^2 and is strictly decreasing in F until reaching a minimum, the constant value of 1.96^2.

The table can be used as follows: 1) Estimate the usual 2SLS (e.g. robust, clustered, etc.) standard error, 2) multiply the standard error by the adjustment factor in the table corresponding to the observed first-stage $\hat{F}$ statistic. This "0.05 tF standard error" can be used for constructing the t-ratio for testing a particular hypothesis, or for constructing 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 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".[16]

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 selected values for significance (confidence) levels of 0.01 (.99), another commonly-used standard in applied research, in Table 3b.

The table shows that the $\frac{\sqrt{c_{.01}(F)}}{2.576}$ function has a similar pattern, but three important differences. First, the adjustment factor now has a vertical asymptote at $F = q_{.99} = 2.576^2$. Second, $c_{.01}(F)$ declines until $F = 252.34$, at which point

F-statistic.

[16] We have also posted code to allow more precise computation of the adjustment factor for any given value of $\hat{F}$

Table 3a: Selected values of tF Standard Error Adjustments, $\frac{\sqrt{c_{.05}(F)}}{1.96}$

F	4.000	4.008	4.015	4.023	4.031	4.040	4.049	4.059	4.068	4.079
$\frac{\sqrt{c_{.05}(F)}}{1.96}$	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
	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
	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
	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
	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
	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
	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
	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
	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	104.68
	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 and the bottom number in each row is the corresponding value of $\frac{\sqrt{c_{.05}(F)}}{1.96}$, where we use $\Phi^{-1}(.975)$ for "1.96". Numerical values in each pair are rounded up (e.g. 4.0051 rounds up to 4.006).

the adjustment factor is 1.059. Finally, we note that $\frac{\sqrt{c_{.01}(F)}}{2.576}$ is uniformly strictly above $\frac{\sqrt{c_{.05}(F)}}{1.96}$. This implies that from a reporting convenience standpoint, one could choose to report only the ".01 tF standard errors" by using the adjustments in Table 3b, and the intervals $\hat{\beta} \pm 2.576 \times$ (".01 tF standard error") and $\hat{\beta} \pm 1.96 \times$ (".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.

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

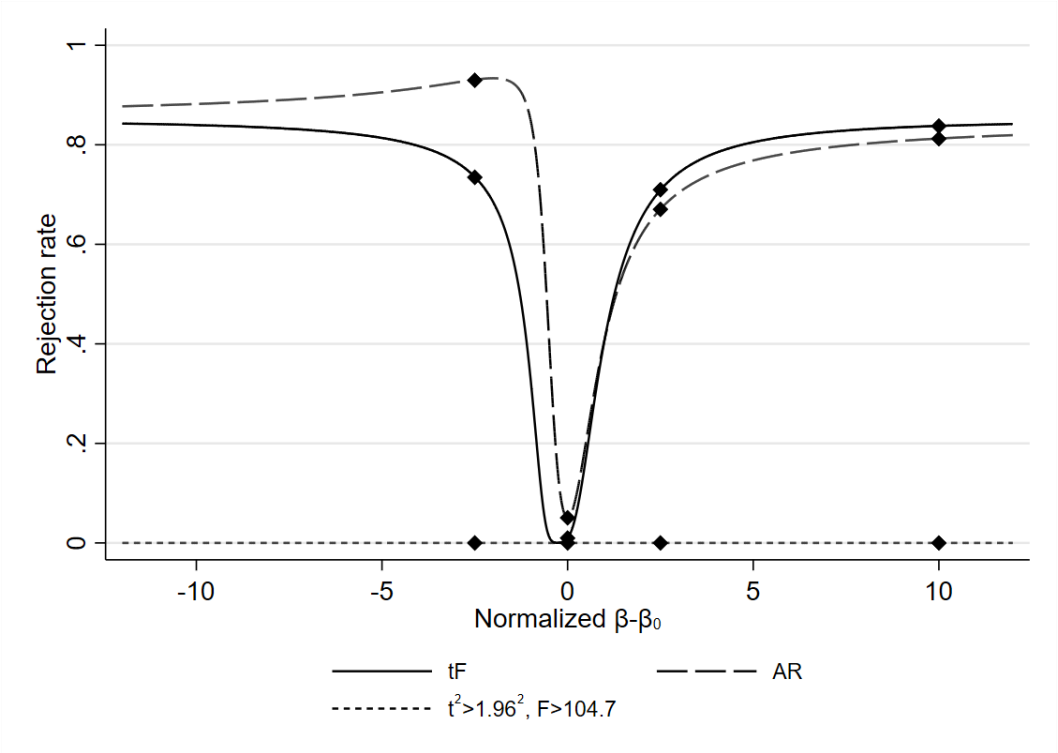
Table 3b: Selected values of tF Standard Error Adjustments, $\frac{\sqrt{c_{.01}(F)}}{2.58}$

F	6.670	6.673	6.676	6.679	6.682	6.685	6.689	6.693	6.697	6.701
$\frac{\sqrt{c_{.01}(F)}}{2.575}$	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
	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
	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
	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
	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
	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
	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
	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
	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	100.07	128.95	174.37	252.34
	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 and the bottom number in each row is the corresponding value of $\frac{\sqrt{c_{.01}(F)}}{2.58}$, where we use $\Phi^{-1}(.995)$ for "2.58". Numerical values in each pair are rounded up (e.g. 6.6712 rounds up to 6.672).

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.

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.

3.3 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 4 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 0.05. We summarize the results below. Note that in our comparisons, we focus only on procedures that allow the research to be completely agnostic about the nuisance

parameters.[17]

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.[18] 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 traditional power curves for the nine combinations given by the three values of $\rho = 0, 0.5, 1$ and the three values of $f_0 = 1, 3, 9$.[19]

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 one does not uniformly dominate the other.[20] 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 *OLS* estimand is upward biased, then the probability of rejecting no effect will be slightly higher for tF than for AR if the true effect is positive.[21] 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.8 includes power curves for the other eight

[17] For example, the approach of Kocherlakota (2020) requires the researcher to assume a lower bound for f_0 for inference.

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

[19] 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.

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

[21] 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 $\beta_0 - \beta$.

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 another.

Given that neither AR nor tF uniformly dominates one another 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 $104.7 > 1.96^2$. [22]

For any realization of the data, the tF and AR confidence sets behave similarly: either both are bounded intervals ($F > q_{1-\alpha}$) or both are unbounded ($F < q_{1-\alpha}$). Then, 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 always finite. We show below that this is true uniformly across all possible values of the

[22] 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* must have infinite unconditional expected length.

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 4 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.[23] Using the same data generating process from 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, and furthermore, there are occasional sharp discontinuities, which is expected from a distribution of lengths with thick tails, which underlies the infinite conditional mean.[24] 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.

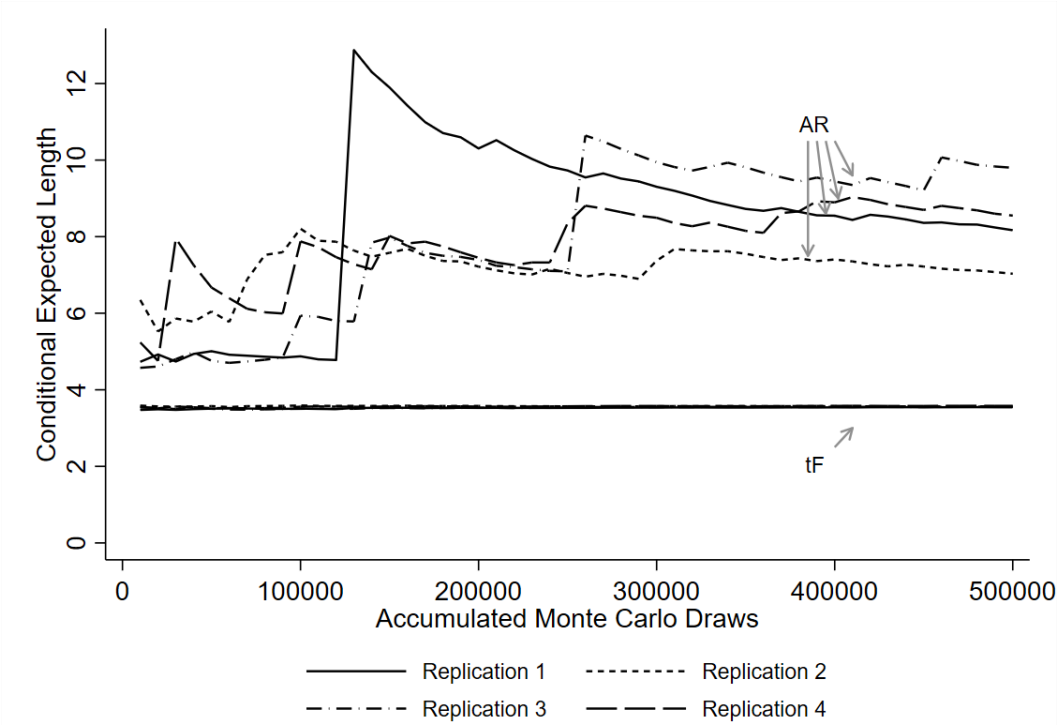
3.4 The tF procedure: Impact on Applications

We now turn to gauging how the tF adjustments to the standard errors would impact practice by 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

[23] 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*).

[24] 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.

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

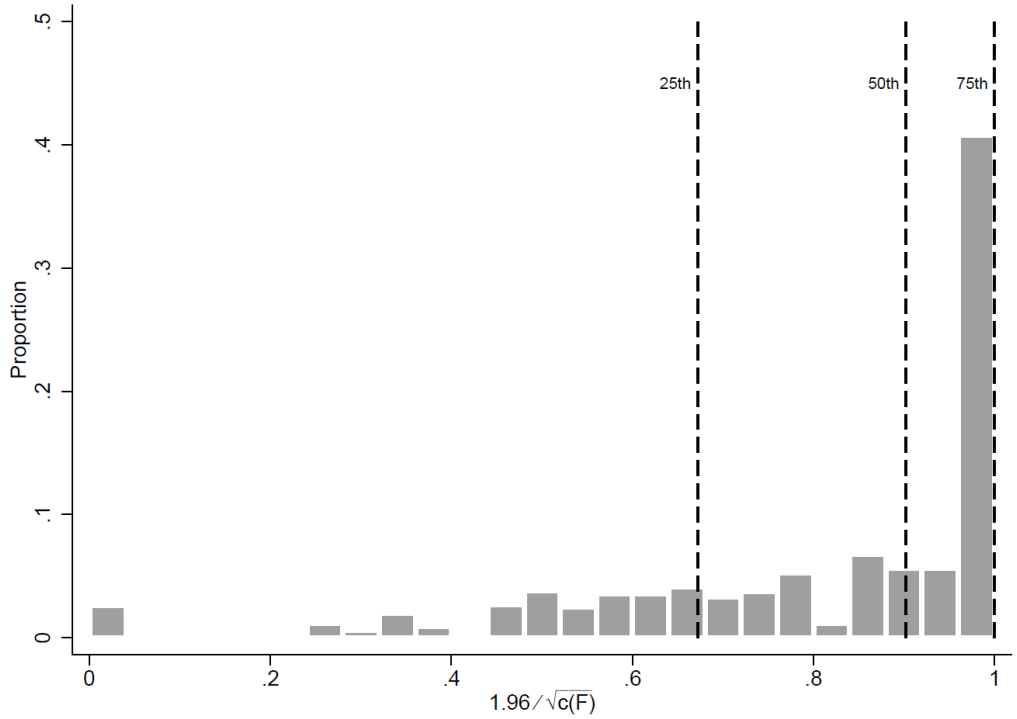


Note: Points on each curve represent the conditional expected length, using the specified number of accumulated Monte Carlo draws, for tF (lower four lines) and AR (upper four lines). Each of the four lines corresponds to an independent set of Monte Carlo draws.

for the reciprocal of the 0.05 tF adjustment factor, which represents the degree to which the reported standard errors are understated. It shows significant mass at values close to 1 (no understatement); the median understatement is about 10 percent while the 25th percentile understatement is about 33 percent. The weighted mean value is 0.801.

Therefore, in this sample of studies, the tF adjustment would be expected to increase confidence intervals by about 25 percent. To understand this magnitude, it is helpful to recall that conventional 95 percent confidence intervals are about 20 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,

Figure 5a: Distribution of $\frac{1.96}{\sqrt{c_{.05}(F)}}$ for AER sample

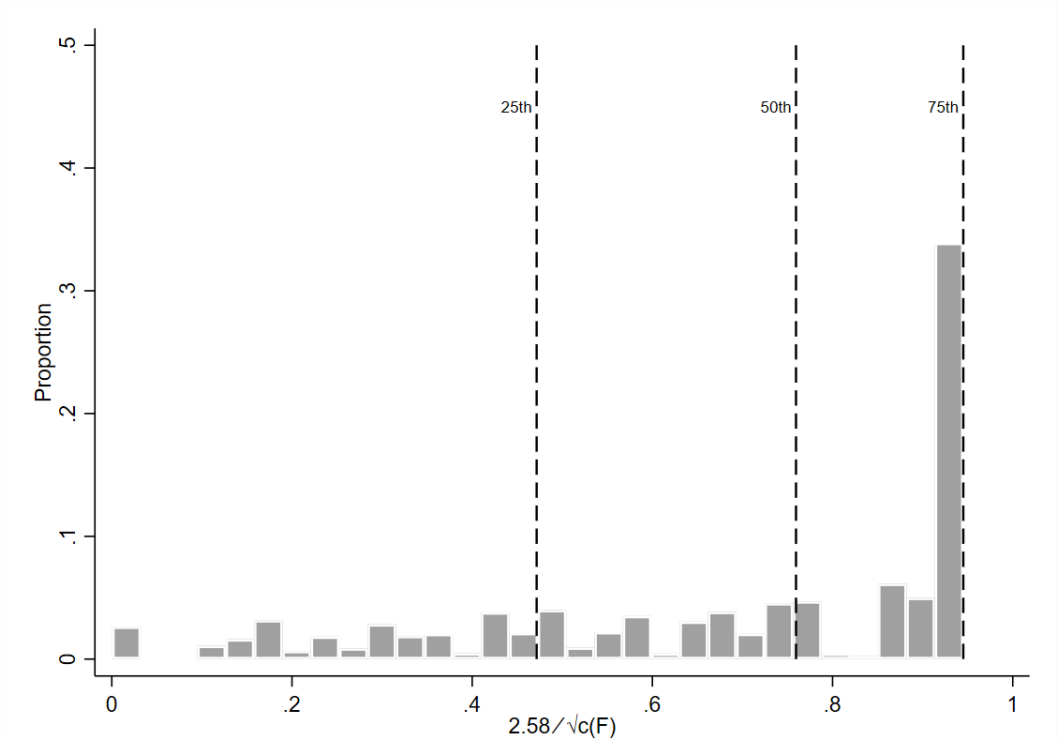


N=847 specifications. The x-axis is the ratio of $\Phi^{-1}(.975)$ to the F-dependent value $\sqrt{c_{.05}(F)}$. All specifications use the derived F statistic, and when not possible, the reported F statistic from the paper. The 6 specifications that report (rounded) first-stage standard errors of zero are excluded. Proportions are weighted; see notes to Table 1. Dashed lines correspond to the (weighted) 25th (0.672), 50th (0.902), and 75th (1.00) percentiles of the distribution.

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, the tF adjustment could be expected to impact standard errors by a magnitude roughly equivalent to erroneously using a 90 percent confidence interval while calling it a 95 percent confidence interval, or using non-clustered standard errors when clustered standard errors are appropriate.

Figure 5b repeats the exercise for the .01 tF adjustments and finds more significant degrees of adjustment, where the weighted median degree of understatement is now about 24 percent, while the weighted mean value is 0.68, implying a 32 percent understatement.

Figure 5b: Distribution of $\frac{2.58}{\sqrt{c_{.01}(F)}}$ for AER sample

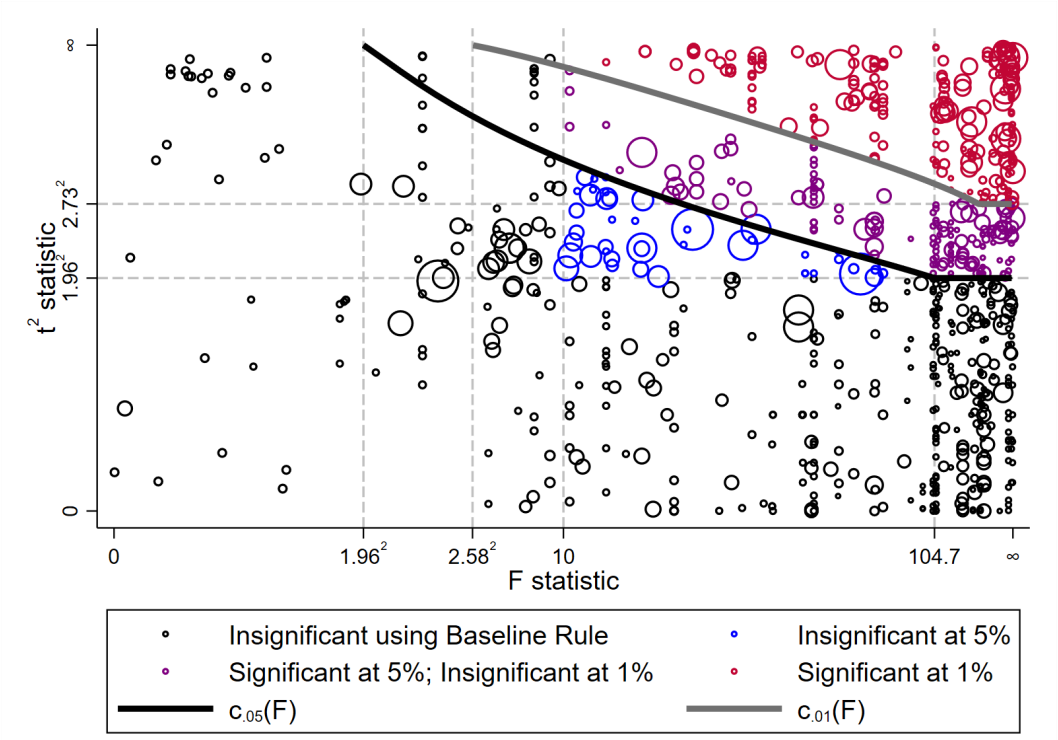


N=847 specifications. The x-axis is the ratio of 2.58 to the F-dependent value $\sqrt{c_{.01}(F)}$. See notes to Figure 5a. Dashed lines correspond to the (weighted) 25th (0.472), 50th (0.759), and 75th (0.945) percentiles of the distribution.

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 $\frac{\frac{t^2}{1.96^2}}{1+\frac{t^2}{1.96^2}}$ and $\frac{\frac{F}{10}}{1+\frac{F}{10}}$ 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. 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 generally viewed the black circles as statistically insignificant estimates by virtue of either the ob-

served t-ratio or the F-statistic.[25] While most of these black circles would remain insignficant using the tF adjustment, at the 5% level, some, by being above the tF critical value function would become significant. The remaining specifications

Figure 6: Statistical Significance in AER sample, using $c_{.05}(F)$ and $c_{.01}(F)$



N=847 specifications. Vertical scale is $\frac{\frac{t^2}{1.96^2}}{1+\frac{t^2}{1.96^2}}$ and horizontal scale is $\frac{\frac{F}{10}}{1+\frac{F}{10}}$. Size of circle is proportional to the weight described in Table 1. Size of the circle is proportional to the weight described in Table 1. The solid black and gray lines are critical value functions $c_{.05}(F)$ and $c_{.01}(F)$, 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 $c_{.05}(F)$. The purple circles represent those that are significant at the 5% but not significant at the 1% level. The red circles represent those that are significant at the 1% level.

(blue, purple, and red circles), under current norms, would most likely have been viewed as statistically significant. Of these, 24% (the blue circles) are in fact statistically insignificant at the 5 percent level, when the tF critical values are applied;

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

the remaining 76% (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 31 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 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).

4 Derivation of Theoretical Results

This section explains how we derive all of the theoretical results discussed in Section 3. Subsection 4.1 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 4.2 defines the tF critical value function, formally states some of its properties, and describes at a high-level the relevant proofs in the Online Appendix. Subsection 4.3 formally states the results on the conditional expected length of the AR and tF confidence sets and describes the proofs. The details of all of the proofs of the results of this Section can be found in the Online Appendix.

4.1 Notation and Preliminaries: Rejection probabilities for t-ratio-based rules

We begin by introducing some additional notation.

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

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

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

where β_0 is a hypothesized value for β, $\hat{t}_{AR}(\beta_0)$ is a "t-ratio form" of the statistic of Anderson and Rubin (1949), so that $\hat{t}^2_{AR}(\beta_0) = AR$. $\hat{V}_N\left(\widehat{\pi\beta}\right), \hat{C}_N\left(\widehat{\pi\beta},\hat{\pi}\right)$, 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 heteroskedastic errors, as well as one-way or two-way clustering, for example. 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

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

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. [26] 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

[26] 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

$$\begin{aligned}
\Pr_{\Delta(\beta_0),\rho,f_0}\left[t^2 > k(F)\right] = \int\int 1\left[t^2(x,y,\rho(\beta_0)) > k\left(y^2\right)\right] \\
\times \varphi\left(x - f_0\frac{\Delta(\beta_0)}{\sqrt{1+2\rho\Delta(\beta_0)+\Delta^2(\beta_0)}}, y - f_0; \rho(\beta_0)\right) dxdy
\end{aligned} \tag{4}$$

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).[27]

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?[28] 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 provides answers to these questions.

[27] 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.

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

4.2 The tF critical value function: Definition and Properties

Equations (2) and (3) allow us to construct the tF critical value function $c_\alpha(F)$, and (4) allows us to verify its ability to control size and to explore its power.

Definition. The tF critical value function $c_\alpha(F)$ is defined as follows:

1. Let Θ denote the set of all pairs $\left(\tilde{F}, \tilde{c}_\alpha\left(F,\tilde{F}\right)\right) \in (q_{1-\alpha},\infty)\times\mathscr{C}^0$ that satisfy the following properties:

   (a) (Piecewise with plateau) Viewed as a function of F, $\tilde{c}_\alpha\left(F,\tilde{F}\right)$ is decreasing in F on $(q_{1-\alpha},\tilde{F}]$ and equal to the constant $\tilde{c}_\alpha\left(\tilde{F},\tilde{F}\right)$ for $F > \tilde{F}$

   (b) (Controls size for $|\rho|=1$, f_0 small) $\Pr_{|\rho|=1,f_0}\left[t^2 > \tilde{c}_\alpha\left(F,\tilde{F}\right)\right] = \alpha$ for $|f_0| < \bar{f}_0$ where $\bar{f}_0 = h\left(\tilde{F},\tilde{c}_\alpha\left(\tilde{F},\tilde{F}\right)\right)$, with $h(x,y) \equiv \frac{x}{\sqrt{y}+\sqrt{x}}$

   (c) ($\tilde{F}$ not too large) $\Pr_{|\rho|=1,f_0=\bar{f}_0}\left[\left\{t^2 > \tilde{c}_\alpha\left(F,\tilde{F}\right)\right\}\cap\left\{F_* < F \le \tilde{F}\right\}\right] = 0$ where

$$F_* \equiv \left[\bar{f}_0 + \Phi^{-1}\left(\Phi\left(\sqrt{\tilde{F}} - \bar{f}_0\right) - (1-\alpha)\right)\right]^2$$

2. Let $c_\alpha(F) \equiv \tilde{c}_\alpha(F,F^*)$, where

$$F^* = \max\left\{\tilde{F}\middle|\left(\tilde{F},\tilde{c}_\alpha\left(F,\tilde{F}\right)\right)\in\Theta \text{ and } \max_{|\rho|\le 1, f_0\neq 0}\Pr\left[t^2 > \tilde{c}_\alpha\left(F,\tilde{F}\right)\right] = \alpha\right\}.$$

The tF critical value function has the structure of a piecewise continuous function with a plateau, as given by property 1(a). The set Θ characterizes a whole class of functions that satisfy a key property of controlling rejection probabilities for small f_0 and $|\rho|=1$ (property 1(b)). By definition, the tF critical value function is given by the element in Θ that gives the smallest plateau, while still controlling size (rejection probabilities across the entire nuisance parameter space, property (2), not just for $|\rho|=1$, and small f_0). For the 5 percent level, $F^* = 104.7$ as labeled in Figure 6 and shown in Table 3a, while for the 1 percent level $F^* = 252.34$ as shown in Table 3b.

We highlight the following points about this critical value function:

1. Property 1(b) is motivated by the numerical analysis of Stock and Yogo (2005) that determined that $\rho = \pm 1$ produced the "worst case" rejection probabili-

ties for a given f_0 under the rule $t^2 > q_{1-\alpha}$. Our analysis based on numerical integration confirms this as well.

2. That Θ is a nonempty set (existence) is not trivial to prove, and the details are in Appendix B.1.

3. Property 1(c) is a technical refinement that excludes from the set Θ those critical value functions such that for $f_0 = \bar{f}_0$, the set of values of F that are in the acceptance region is a non-convex set. This simplifies the definition and derivation of the tF critical value function.[29]

4. It should be clear that, by definition, the level of the plateau of $c_\alpha(F)$ cannot be lowered further without violating one of the properties. In Appendix B.2, we show that any critical value function that is somewhere strictly below $c_\alpha(F)$ will over-reject for some f_0, ρ. It is in this sense that the acceptance region for $c_\alpha(F)$ in t^2, F space is as "small" as possible.

Deferring details to Appendix B.1, we now sketch the key elements of constructing the tF critical value function. It proceeds in the following steps:

1. Solve the following functional equation for $\kappa(F)$:

$$\kappa\left((g)^2\right) = \frac{g^2(g-h)^2}{h^2}, \text{ where } h = \frac{F}{\sqrt{\kappa(F)} + \sqrt{F}} \text{ and}$$
$$g = h + \Phi^{-1}\left(\Phi\left(\sqrt{F} - h\right) - (1-\alpha)\right).$$

This functional equation stems from considering small f_0 (without loss of generality, > 0) and $\rho = 1$.[30] The key point is that when $\rho = 1$, the square of

[29] It is possible to relax this restriction and therefore enlarge the set Θ. Doing so has no impact on the 5 percent level critical values because property 2 is the binding constraint for determining F^* (which is equal to 104.7). Our numerical analysis of the 1 percent level, suggests that removing property 1(c) would mean that the $\tilde{F}$ value for the elements of Θ would be unbounded: the critical value function for the 1 percent would asymptote from above to 2.58^2. Property 1(c) ensures the existence of a maximum value F^*. For details on the algorithm to consider crticial value functions that violate Property 1(c), see Lee et al. (2020).

[30] Following the same logic considering $f_0 < 0$ and/or $\rho = -1$ results in the same function.

the t-ratio, for a fixed value of f_0, is a deterministic function of f (a quartic polynomial in f)

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

which means that the rejection probability involves the set of f such that this function of f is above the critical value function; recall that $f - f_0$ is a standard normal random variable. $\kappa(F)$ will be used as the decreasing part of a candidate critical value function $\tilde{c}_\alpha\left(F, \tilde{F}\right)$.

2. Any two solutions to the functional equation will coincident in a neighborhood of $F = q_{1-\alpha}$, with the only difference being the length of the interval over which κ is defined. Find the candidate function $\tilde{c}_\alpha\left(F, \tilde{F}\right)$ with the largest $\tilde{F}$ such that it satisfies Properties 1(a), 1(b), 1(c) and controls size (using equation (4)) as described in Property 2.

In Appendix B.1, we show that our functional equation can be reformulated in terms of a mapping that satisfies the conditions for application of the existence and uniqueness theorem of Fefferman (2021), regarding invariant curves for degenerate hyperbolic maps in the plane.

Finally, we note that while we are both motivated by, and leverage, the conclusions of the numerical analysis of Stock and Yogo (2005) that $\rho = \pm 1$ delivers the worst-case rejection probabilities, we provide additional analysis that further corroborates their finding. First, we use numerical integration of the expression in (4) to compute the rejection probabilities under the null.[31] Thus, even if there were some regions where rejection probabilities were larger for $|\rho| < 1$, they would be limited to that which could not be detected given the precision of numerical integration. Second, we present a theoretical result that establishes that $\rho = \pm 1$ represents worst case rejection probabilities in a particular "corner" of the nuisance parameter space. We state the results below, deferring the proofs to Appendices B.3.

Size control for small f_0 with $|\rho|$ in a neighborhood of 1. *Under the null, for any $|\rho|$ arbitrarily close to 1, there exists $\bar{f}_0$ for which $\Pr_{f_0,\rho}\left[t^2 > c_\alpha(F)\right] < \alpha$ for all $f_0 < \bar{f}_0$.* By construction, $c_\alpha(F)$ controls rejection probabilites for the case of

[31] Stock and Yogo (2005) use Monte Carlo integration to evaluate the expression

$\rho = \pm 1$ for small f_0. The above result says that it also controls rejection probabilites for $|\rho|$ close to 1 for sufficiently small f_0.

4.3 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 3.3, 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. We find that *tF* intervals can be expected to be shorter. Indeed, the condi-

tional expected length for the *AR* confidence interval is infinite, while the expected length of the tF interval is finite.

We provide some intuition for this result. Appendix C.1 shows that, conditional on $F > q_{1-\alpha}$ the three following confidence interval lengths for *IV* $\left(\hat{L}_{IV}\right)$, *AR* $\left(\hat{L}_{AR}\right)$, and tF $\left(\hat{L}_{tF}\right)$ converge in distribution under weak-IV asymptotics as follows:

$$\hat{L}_{IV} \xrightarrow{d} L_{IV} \equiv 2\sqrt{q_{1-\alpha}}\sqrt{1-2\rho\frac{t_{AR}(\beta)}{f}+\frac{t_{AR}^2(\beta)}{f^2}}\frac{1}{\sqrt{F}}\sqrt{V_\Omega}, \tag{5}$$
$$\hat{L}_{AR} \xrightarrow{d} L_{AR} \equiv \frac{\sqrt{F}\sqrt{F-q_{1-\alpha}\left(1-\tilde{\rho}^2\right)}}{F-q_{1-\alpha}}L_{IV} \text{ and } \hat{L}_{tF} \xrightarrow{d} L_{tF} \equiv \frac{\sqrt{c_\alpha(F)}}{\sqrt{q_{1-\alpha}}}L_{IV}.$$

where

$$\tilde{\rho}^2 = \frac{\left(-t_{AR}(\beta)+\rho f\right)^2}{\left(f^2-2\rho t_{AR}(\beta) f+t_{AR}^2(\beta)\right)} \text{ and } V_\Omega = \frac{AV\left(\widehat{\pi\beta}\right)-2\beta AC\left(\widehat{\pi\beta},\hat{\pi}\right)+\beta^2 AV\left(\hat{\pi}\right)}{AV\left(\hat{\pi}\right)}.$$

The length of the *IV* $(1-\alpha)$ confidence interval, $\hat{L}_{IV}$ converges in distribution to a function of two jointly normal random variables $t_{AR}(\beta), f$ and the constants ρ, V_Ω, and $q_{1-\alpha}$. Meanwhile, the limiting distribution of $\hat{L}_{AR}$ is equal to the random variable L_{IV} times an inflation factor given by $\frac{\sqrt{F}\sqrt{F-q_{1-\alpha}\left(1-\tilde{\rho}^2\right)}}{F-q_{1-\alpha}}$, which itself is a function of the random variables $t_{AR}(\beta), f$ and constants $q_{1-\alpha}$ and ρ. Finally, the tF inflation factor $\sqrt{\frac{c_\alpha(F)}{q_{1-\alpha}}}$ is only a function of the random variable F.

As discussed in Andrews, Stock and Sun (2019) and in Mikusheva (2010), it should be noted that, strictly speaking, L_{AR} is not always the length of the confidence interval. It is the length of a bounded interval that is equal to the confidence interval when $F > q_{1-\alpha}$, but it is the *complement* of the confidence set when $F < q_{1-\alpha}$. That is, when $F < q_{1-\alpha}$ the *AR* confidence set is non-convex: it covers the real line *except* for the bounded interval that has length L_{AR}.

In Appendices C.2 and C.3, we show that for all $\rho, f_0 \neq 0$

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

The intuition behind these results can be seen from Equation (5). Whenever $F > q_{1-\alpha}$, it is clear from the inflation factors for both L_{AR} and L_{tF}, that the lengths of the intervals asymptote to infinity as F approaches $q_{1-\alpha}$. It turns out that the inflation factor essentially explodes too fast for the conditional expectation to exist for L_{AR}. As for L_{tF}, the finiteness of $E[L_{tF}|F > q_{1-\alpha}]$ may not be surprising in light of our finding in Appendix B.1 that

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

which implies, given the continuity of $c_\alpha(F)$, that $\sqrt{c_\alpha(F)} \leq \sqrt{\frac{K}{F-q_{1-\alpha}}}$ in a neighborhood to the right of $F = q_{1-\alpha}$ for some finite K, which implies integrability of $\sqrt{c_\alpha(F)}$ in that small neighborhood – the function $\sqrt{c_\alpha(F)}$ does not explode "too fast" as F tends to $q_{1-\alpha}$.

While tF confidence sets are expected to be (infinitely) shorter when $F > q_{1-\alpha}$, AR confidence sets have smaller expected lengths when $F < q_{1-\alpha}$. For this latter case, the tF confidence set consists of the entire line, but the AR confidence set is either the real line or the union of two unbounded intervals. 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.[32] If, for example, in practice researchers effectively ignore the non-convexity and simply use the whole real line as the confidence set, then it would no longer be "shorter" when $F < q_{1-\alpha}$. In other words, the expected difference in lengths between a "convexified" AR confidence set and the tF interval would always favor tF.

5 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, and the work of Staiger and

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

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.

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(6): 2453–2483.

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

Fefferman, Charles. 2021. "Invariant Curves for Degenerate Hyperbolic Maps of the Plane."

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(2): 467–475.

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(2): 281–355.

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

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.

Appendix: For Online Publication

Click here for current version of Online Appendix

http://www.princeton.edu/∼davidlee/wp/SupplementarytF.html