Valid $t$-Ratio Inference for IV

David S. Lee  
*Princeton University*

Justin McCrary  
*Columbia Law School, jmccrary@law.columbia.edu*

Marcelo J. Moreira  
*FGV EPGE*

Jack Porter  
*University of Wisconsin*

Follow this and additional works at: https://scholarship.law.columbia.edu/faculty_scholarship

Part of the Law and Economics Commons

**Recommended Citation**

Available at: https://scholarship.law.columbia.edu/faculty_scholarship/3902

This Article is brought to you for free and open access by the Faculty Publications at Scholarship Archive. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of Scholarship Archive. For more information, please contact scholarshiparchive@law.columbia.edu.
Valid \( t \)-Ratio Inference for IV

By David S. Lee, Justin McCrary, Marcelo J. Moreira, and Jack Porter*†

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 \( t_F \) 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. \( t_F \) confidence intervals have shorter expected length than those of Anderson and Rubin (1949), whenever both are bounded. (JEL C13, C26)

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

\[
Y = \beta X + u, \quad \text{where} \ C(u, Z) = 0, \ C(Z, X) \neq 0.
\]

Conducting hypothesis tests and constructing confidence sets for \( \beta \) 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

*Lee: Princeton University and NBER (email: davidlee@princeton.edu); McCrary: Columbia University and NBER (email: jmccrary@law.columbia.edu); Moreira: FGV EPGE (email: mjmoreira@fgv.br); Porter: University of Wisconsin (email: jporter@ssc.wisc.edu). Isaiah Andrews was the coeditor for this article. 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 Josh Angrist 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 Tailey, 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, 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. For supplementary material, including updates to the original online Appendix and a STATA package to compute \( t_F \) critical values/standard error adjustments, please visit https://www.princeton.edu/~davidlee/wp/SupplementarytF.html.

†Go to https://doi.org/10.1257/aer.20211063 to visit the article page for additional materials and author disclosure statements.

1It 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.
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. To deal with this problem, researchers have relied upon the first-stage $F$-statistic as a pretest 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-identified case—a common case for applied work, and the exclusive focus of the current paper.

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, pretesting 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 $\hat{\beta} \pm 1.96 \times \hat{\text{se}}(\hat{\beta})$ as a “95 percent confidence interval” (after pretesting 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 percent confidence interval.

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 pretesting 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 by Moreira and Moreira (2019).

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).
in tables below for 95 percent and 99 percent 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 percent 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 percent 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, 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 reassess 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 comments, replies, and AEA Papers and Proceedings, 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 (i) outcome, (ii) endogenous regressor, (iii) instrument, and (iv) combination of covariates. The dataset contains 1,311 specifications from 61 studies; among those studies, the average number of specifications is 21.5, with a median of 9, and with twenty-fifth and seventy-fifth 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 are reported: the coefficients and standard errors from (i) only the 2SLS, (ii) the 2SLS and first-stage regression, (iii) the 2SLS and the reduced-form regression of the outcome on the instrument, and (iv) 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 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).

\(^8\) 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 the online Appendix.


\(^10\) 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.
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.\footnote{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 $\hat{F}$ 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 displays the histogram of the $F$-statistics in our sample on a logarithmic scale. The weighted twenty-fifth percentile, median, and seventy-fifth 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.\footnote{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.} 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 2SLS

<table>
<thead>
<tr>
<th>Combinations of regressions reported</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-stage least squares</td>
<td>445</td>
<td>132</td>
<td>577</td>
</tr>
<tr>
<td>(0.339)</td>
<td>(0.101)</td>
<td>(0.44)</td>
<td></td>
</tr>
<tr>
<td>[0.251]</td>
<td>[0.088]</td>
<td>[0.339]</td>
<td></td>
</tr>
<tr>
<td>Two-stage least squares and first stage</td>
<td>247</td>
<td>212</td>
<td>459</td>
</tr>
<tr>
<td>(0.188)</td>
<td>(0.162)</td>
<td>(0.35)</td>
<td></td>
</tr>
<tr>
<td>[0.204]</td>
<td>[0.154]</td>
<td>[0.358]</td>
<td></td>
</tr>
<tr>
<td>Two-stage least squares and reduced form</td>
<td>13</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>(0.01)</td>
<td>(0.005)</td>
<td>(0.015)</td>
<td></td>
</tr>
<tr>
<td>[0.024]</td>
<td>[0.035]</td>
<td>[0.059]</td>
<td></td>
</tr>
<tr>
<td>Two-stage least squares, first stage, and reduced form</td>
<td>181</td>
<td>74</td>
<td>255</td>
</tr>
<tr>
<td>(0.138)</td>
<td>(0.056)</td>
<td>(0.195)</td>
<td></td>
</tr>
<tr>
<td>[0.15]</td>
<td>[0.094]</td>
<td>[0.244]</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>886</td>
<td>425</td>
<td>1,311</td>
</tr>
<tr>
<td>(0.676)</td>
<td>(0.324)</td>
<td>(1)</td>
<td></td>
</tr>
<tr>
<td>[0.628]</td>
<td>[0.372]</td>
<td>[1]</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Number of specifications: 1,311. 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.
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 $AR$ confidence 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) and Keane and Neal (2021). 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.

This known and quantified distortion is eliminated by $tF$ inference, which takes as given the common practice of computing the $2SLS$ and standard errors, and provides 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 application of the $tF$ adjustments impacts the statistical inferences in our sample of $AER$ studies.

$^{13}$ We focus on 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.
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 \( Y, X, Z \).

The key statistics are given by

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

where \( \hat{\beta} \) is the instrumental variable estimator, \( \hat{V}_N(\hat{\beta}) \) 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 \( \beta \) and the (possibly false) hypothesized value \( \beta_0 \). \( \hat{f} \) 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 \( \hat{t}^2 \overset{d}{\rightarrow} t^2 \). 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 \( \pm 1.96 \) for testing hypotheses at the 5 percent level. More generally, the critical values \( \pm \sqrt{q_{1-\alpha}} \) are used for tests at the \( \alpha \) 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 (i) 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 (ii) 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 nonnormal distribution, which in some cases might be “close” to the standard normal, but in other cases, the deviation from normality can be significant.

\(^{14}\) 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.
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 rearrangement of terms, the two normal variables can be seen as \( f \) and \( t_{AR} \), where \( \hat{f} \sim d \rightarrow f \) and \( f \) has mean \( f_0 \equiv \frac{\pi}{\sqrt{N} \text{AV}(\hat{\pi})} \) and unit variance, where \( \text{AV}(\hat{\pi}) \) is the asymptotic variance of \( \hat{\pi} \) and \( t_{AR} \) is a standard normal with \( AR = t^2_{AR} \). The correlation \( \rho \) of \( f \) and \( t_{AR} \) is the correlation of \( Zu \) and \( Zv \).  

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 \( \rho \), where \( E[F] = f_0^2 + 1 \). The figure illustrates that with low values of \( \rho \) (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 \( \rho \) 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, \( \rho \) 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 \( \rho \) or \( E[F] \), Staiger and Stock (1997) and Stock and Yogo (2005) propose to use the observed first-stage \( \hat{F} \). Reexpressing 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

\[
\text{Pr}[t^2 > c^*, F > F^*] \leq \alpha
\]

for a prespecified significance level \( \alpha \). 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

\[
\frac{\hat{f}}{\sqrt{N} \text{AV}(\hat{\pi})}
\]

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

16 As we explain in detail in Section III, rejection probabilities displayed in panel A of Figure 2 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 Staiger and Stock (1997, Section 4B). 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.
procedure that sets the confidence interval to the entire real line if $F < F^*$, and otherwise uses $\pm \sqrt{c^*} \times \text{(standard error)}$ 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 prespecified significance level $\alpha$, where $c_\alpha(F)$ is a smooth function of $F$, instead
of a step function.\footnote{As we will show below, inference based on $t_F$ has significant power advantages over inference based on a test that uses constant thresholds $c^*, F^*$; furthermore, $t_F$ confidence intervals will have shorter expected length compared to that of $AR$ when both are bounded intervals.} As we will show below, inference based on $t_F$ has significant power advantages over inference based on a test that uses constant thresholds $c^*, F^*$; furthermore, $t_F$ confidence intervals will have shorter expected length compared to that of $AR$ when both are bounded intervals.

**B. The $t_F$ Procedure: Critical Values and Valid Inference**

Table 3A 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\footnote{Similar in spirit to the Bonferroni approach discussed in Staiger and Stock (1997, Section IVB), the probability considered is an unconditional one. See Chioda and Jansson (2005) for an analysis of inference conditional on the observed $F$-statistic.} 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 normalizes 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: (i) estimate the usual 2SLS (e.g., robust, clustered, etc.) standard error, (ii) 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 $t_F$ standard error,” and can be used for constructing the $t$-ratio for testing a particular hypothesis, or for constructing 95 percent confidence intervals using $\beta \pm 1.96 \times \text{"0.05 tF standard error"}$. 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

$$1.727 + \frac{10.253 - 10}{10.253 - 9.835} \times (1.767 - 1.727) = 1.751$$


$\sqrt{c_{0.01}(F)}$ tends to infinity as $F$ tends to 2.576\footnote{We have also posted STATA code at https://www.princeton.edu/~davidlee/wp/SupplementarytF.html to allow more precise computation of the adjustment factor for any given value of $\hat{F}$.} from above, and it is strictly decreasing in $F$ until reaching a minimum, the constant value of 2.576, when $F$ reaches around 169.2. The third line of numbers in each row normalizes the critical values by 2.576, and therefore represent a standard error adjustment factor. Adjusted standard errors can be constructed using the table as follows: (i) estimate the usual 2SLS (e.g., robust, clustered, etc.) standard error, (ii) 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 $t_F$ standard error,” and can be used for constructing the $t$-ratio for testing a particular hypothesis, or for constructing 95 percent confidence intervals using $\beta \pm 1.96 \times \text{"0.05 tF standard error"}$. 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

$$1.727 + \frac{10.253 - 10}{10.253 - 9.835} \times (1.767 - 1.727) = 1.751$$


The table shows that the $\sqrt{c_{0.01}(F)}$ function has a similar pattern, but with 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 $\sqrt{c_{0.01}(F)}$ 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 $t_F$ standard errors” by using the adjustments in Table 3B and the intervals $\hat{\beta} \pm 2.576 \times \text{"0.01 tF standard error"}$ and $\hat{\beta} \pm 1.96 \times \text{"0.01 tF standard error"}$ would be assured of confidence levels at
both the ninety-ninth and ninety-fifth percent levels. The cost for this reporting convenience is that the latter interval would be unnecessarily conservative.

We verify that the $t_F$ 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 $t_F$ procedure for the same values of $\rho$ 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.

C. The $t_F$ 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 $t_F$ 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

### Table 3A

Panel A. Selected values of $t_F$ critical values, $\sqrt{c_{0.05}(F)}$, and $t_F$ standard error adjustments, $\sqrt{c_{0.05}(F)/1.96}$

<table>
<thead>
<tr>
<th>$F$</th>
<th>$\sqrt{c_{0.05}(F)}$</th>
<th>$\sqrt{c_{0.05}(F)/1.96}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.000</td>
<td>4.008</td>
<td>4.015</td>
</tr>
<tr>
<td>18.656</td>
<td>18.236</td>
<td>17.826</td>
</tr>
<tr>
<td>7.519</td>
<td>7.905</td>
<td>8.891</td>
</tr>
<tr>
<td>4.090</td>
<td>4.101</td>
<td>4.113</td>
</tr>
<tr>
<td>14.859</td>
<td>14.524</td>
<td>14.197</td>
</tr>
<tr>
<td>7.811</td>
<td>7.244</td>
<td>6.921</td>
</tr>
<tr>
<td>4.229</td>
<td>4.247</td>
<td>4.265</td>
</tr>
<tr>
<td>11.834</td>
<td>11.568</td>
<td>11.308</td>
</tr>
<tr>
<td>6.038</td>
<td>5.902</td>
<td>5.770</td>
</tr>
<tr>
<td>4.449</td>
<td>4.477</td>
<td>4.507</td>
</tr>
<tr>
<td>4.809</td>
<td>4.701</td>
<td>4.595</td>
</tr>
<tr>
<td>4.803</td>
<td>4.849</td>
<td>4.897</td>
</tr>
<tr>
<td>7.507</td>
<td>7.338</td>
<td>7.173</td>
</tr>
<tr>
<td>3.830</td>
<td>3.744</td>
<td>3.660</td>
</tr>
<tr>
<td>5.393</td>
<td>5.472</td>
<td>5.556</td>
</tr>
<tr>
<td>5.979</td>
<td>5.844</td>
<td>5.713</td>
</tr>
<tr>
<td>3.051</td>
<td>2.982</td>
<td>2.915</td>
</tr>
<tr>
<td>6.440</td>
<td>6.585</td>
<td>6.741</td>
</tr>
<tr>
<td>4.762</td>
<td>4.655</td>
<td>4.550</td>
</tr>
<tr>
<td>2.430</td>
<td>2.375</td>
<td>2.322</td>
</tr>
<tr>
<td>8.473</td>
<td>8.773</td>
<td>9.098</td>
</tr>
<tr>
<td>3.793</td>
<td>3.707</td>
<td>3.624</td>
</tr>
<tr>
<td>1.935</td>
<td>1.892</td>
<td>1.849</td>
</tr>
<tr>
<td>13.048</td>
<td>13.796</td>
<td>14.631</td>
</tr>
<tr>
<td>3.021</td>
<td>2.953</td>
<td>2.886</td>
</tr>
<tr>
<td>1.542</td>
<td>1.502</td>
<td>1.407</td>
</tr>
<tr>
<td>27.058</td>
<td>29.967</td>
<td>33.457</td>
</tr>
<tr>
<td>2.406</td>
<td>2.352</td>
<td>2.299</td>
</tr>
<tr>
<td>1.228</td>
<td>1.200</td>
<td>1.173</td>
</tr>
</tbody>
</table>

Notes: 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).
\[ c^* = 1.96^2 \]
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 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 \( \beta \) is the true parameter, while \( \beta_0 \) 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, \( \rho \) and \( f_0 \). Any combination of \( \rho \) 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 \( f_0 = 1, 3, 9 \).\(^{22}\)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sqrt{c_{0.01}(F)} )</td>
<td>91.097</td>
<td>87.924</td>
<td>84.862</td>
<td>81.907</td>
<td>79.054</td>
<td>76.301</td>
<td>73.644</td>
<td>71.079</td>
<td>68.604</td>
<td>66.214</td>
</tr>
<tr>
<td>( \sqrt{c_{0.01}(F)/2.576} )</td>
<td>35.366</td>
<td>34.135</td>
<td>32.946</td>
<td>31.798</td>
<td>30.691</td>
<td>29.622</td>
<td>28.591</td>
<td>27.595</td>
<td>26.634</td>
<td>25.706</td>
</tr>
</tbody>
</table>

\(^{20}\) 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 \( \sqrt{\frac{E[Z^2]}{\sqrt{\text{var}^2}}} \).

\(^{22}\) 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
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. In particular, when the alternative value of $\beta$ is sufficiently larger than $\beta_0$, then $tF$ becomes slightly more powerful, while the opposite is true when $\beta$ is smaller than $\beta_0$. 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 estimated 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. 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.

Online Appendix Figure A2 includes power curves for the other eight scenarios for $\rho, f_0$. The pattern of results mirror those described above, with the additional observations that (i) the power curves for $AR$ are consistently higher for $\rho = 0$, and (ii) the differences between $tF$ and $AR$ (for any $\rho$) are negligible with $f_0 = 9$, but (iii) 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.

Note: A black diamond represents the rejection probability, from 250,000 Monte Carlo simulations, each with a sample size of 1,000.
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 unconditioned expected length. Conditional on the event \(F > 1.96^2\), it is immediately clear that the step rule of \(c^* = 1.96, F^* = 104.7\) will also have infinite expectation since \(104.7 > 1.96^2\).

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 online Appendices C.2 and C.3. Here, we provide a simple visual of this result via a Monte Carlo exercise, shown in Figure 4. 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. 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. 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.

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 (2009a), and discussed in Angrist and Pischke (2009b).

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.
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 \( \sqrt{\frac{c_0(F)}{q_{1-\alpha}}} \) (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. It shows significant mass at values close to 1 (no understatement); the median reciprocal is 0.902 (understated by about 10 percent) while the twenty-fifth 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 sample of studies suggests that for one-quarter of specifications, the tF adjustment would increase confidence intervals, at a minimum, by a factor of \( \frac{1}{0.672} \approx 1.49 \), 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 nonclustered standard errors and found that the clustered standard errors were understated by about 20 percent.

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

\[\text{\footnote{We focus on the reciprocal because the adjustment factor itself has some very large numbers. For any given study, we know that its true average will be infinite because there will always be some positive probability that F < q_{1-\alpha}.}}\]
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 interval, or substantially more severe than using nonclustered 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 $t^2/1.96^2$ and $F/10 \sqrt{1+F/10}$ 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 $t^2$

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.
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 observed $t$-ratio or the $F$-statistic. While most of these black circles would remain insignificant using the $tF$ adjustment, at the 5 percent level, some, by being above the $tF$ critical value function would become significant.

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

---

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.
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 $tF$).

E. A Priori Restrictions on $\rho$

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 $\rho$. While the particular values $|\rho| = 1$ are useful in derivations to provide a worst case, valid inference applies to all values of $\rho$ between $-1$ and 1. Thus, for the just-identified IV model, being agnostic about $E[F]$ and $\rho$ is a requirement for practitioners who wish to rely solely on the textbook IV assumptions that $C(Z,u) = 0$ and $C(Z,X) \neq 0$.

---

31 Our setting allows for heteroskedastic, clustered, and/or autocorrelated errors. Nevertheless, the parameter $\rho$ simplifies to the usual endogeneity coefficient $\text{Corr}(Y - X\beta, X - \pi Z)$ which practitioners have in mind if errors are (conditionally on $Z$) independent and homoskedastic.
Adding restrictions beyond the textbook IV assumptions, for example, with a priori information on the parameter $\rho$, is possible. As referenced in Subsection IIIA, one could ask, “What additional assumption about $\rho$ could be imposed on the data generating process to allow the $\pm 1.96$ 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 $\rho$ 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 $\rho$ 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 $\pm 1.96$ critical values, and assuming that $|\rho| \leq 0.76$ corresponds to a 10 percent level of significance.33 In online Appendix A.8.1, we provide the necessary inflation factors to the $\pm 1.96$ 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 $\rho$ one might expect to see in practice. It is of course impossible to make definitive quantitative statements about the true magnitude of $\rho$ or $\beta$, since they are both unknown parameters; also a full meta-analysis is beyond the scope of this paper. Nevertheless, as discussed in online Appendix A.8.3, it is possible to use data to obtain a valid confidence set on $\rho$. The data from our sample of AER studies show that (i) the confidence intervals for $\rho$ include a broad range of values, with 24 percent of the specifications including values as large as 0.9 in absolute value, and that (ii) in 18 percent of the specifications, the data would have rejected the hypothesis that $|\rho| \leq 0.565$. Online Appendix A.8.2 also points out that assuming $|\rho| \leq 0.565$ is equivalent to placing bounds on $\beta$.34 For this same sample, about 30 percent of the time, assuming $|\rho| < 0.565$ is tantamount to assuming a priori that $\beta$ is not equal to zero.

Whether or not one explores specific restrictions on $\rho$, 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 $|\rho| \leq 0.565$.

---

32 Note that the necessary bound on $|\rho|$ depends on the desired significance/confidence level. For example, if $1/99$ percent significance/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.

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. Derivation of Theoretical Results

This section explains how we derive all of the theoretical results discussed in Section II. Subsection IIIA 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 IIIB defines the \( tF \) critical value function, formally states some of its properties, and describes relevant proofs. Subsection IIIC 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.

A. Notation and Preliminaries: Rejection Probabilities for \( t \)-Ratio-Based Rules

We begin by introducing some additional notation. Define

\[
\hat{t}_{AR}(\beta_0) \equiv \frac{\hat{\pi}(\hat{\beta} - \beta_0)}{\hat{se}(\hat{\pi}(\hat{\beta} - \beta_0))} = \frac{\hat{\pi}(\hat{\beta} - \beta_0)}{\sqrt{\hat{V}_N(\hat{\pi}\beta) - 2\beta_0 \hat{C}_N(\hat{\pi}\beta, \hat{\pi}) + \beta_0^2 \hat{V}_N(\hat{\pi})}},
\]

\[
\hat{u}_0 = (Y - X\beta_0) - Z\hat{\pi}(\hat{\beta} - \beta_0),
\]

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

where \( \beta_0 \) is a hypothesized value for \( \beta \) and \( \hat{t}_{AR}(\beta_0) \) is a “\( t \)-ratio form” of the statistic of Anderson and Rubin (1949), so that \( \hat{t}_{AR}^2(\beta_0) = AR \); see, for example, Feir, Lemieux, Marmer (2016), Lee, McCrary, Moreira, and Porter (2020), and Van de Sijpe and Windmeijer (2021). \( \hat{V}_N(\hat{\pi}\beta), \hat{C}_N(\hat{\pi}\beta, \hat{\pi}), \) and \( \hat{V}_N(\hat{\pi}) \) are elements of the estimator for the variance-covariance matrix of the reduced form and first-stage estimators \( \hat{\pi}\beta \) 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 online 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 error structures and their corresponding heteroskedastic and autocorrelation consistent (HAC) variance estimators (including one- or two-way clustering). As examples of this notation, if we consider the homoskedastic case, \( \hat{\rho}(\beta_0) \) 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

\[
\hat{t}^2 = \frac{\hat{t}_{AR}^2(\beta_0)}{1 - 2\hat{\rho}(\beta_0)\hat{t}_{AR}(\beta_0) + \hat{t}_{AR}^2(\beta_0)}.
\]
which is a numerical equivalence that can be shown using the definitions above and with some rearrangement of terms, as shown in online 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

\[ t^2 \overset{d}{\to} t^2(t_{AR}(\beta_0), f, \rho(\beta_0)) \equiv \frac{t^2_{AR}(\beta_0)}{1 - 2\rho(\beta_0) \frac{t_{AR}(\beta_0)}{f} + \frac{t^2_{AR}(\beta_0)}{f^2}}, \]

where

\[ (t_{AR}(\beta_0)) \sim N\left( \frac{\Delta(\beta_0)}{f}, \frac{\Delta(\beta_0)}{f} \right) \cdot \left( 1, \frac{\rho(\beta_0)}{1} \right), \]

\[ \Delta(\beta_0) = \frac{\sqrt{V(Zv)}}{\sqrt{V(Zu)}}(\beta - \beta_0), \quad \text{and} \quad \rho(\beta_0) = \frac{\rho + \Delta(\beta_0)}{\sqrt{1 + 2\rho \Delta(\beta_0) + \Delta^2(\beta_0)}}, \]

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

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 general critical value function that could depend on \( F \):

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

\[35\] 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.
In all cases, the rejection probability can be expressed as

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

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 (i) illustrate the magnitude of inferential distortions caused by the usual \(t\)-ratio procedure (Figure 2 panel A), (ii) verify that the \(tF\) critical value function controls the significance level, as intended (Figure 2 panel B), and (iii) construct power functions (Figure 3 and online Appendix Figure A2).\(^{36}\)

**Remark:** In addition, expression (4) also allows us to answer the following questions: (i) What restrictions on the nuisance parameter space \(f_0,\rho\) could one impose so that the usual \(t\)-ratio procedure has the intended significance level?\(^{37}\) (ii) For single threshold rules, what minimal threshold for \(F^*\) could one use if \(c^*\) is set to the nominal value \(q_{1-\alpha}\)? and (iii) How do these answers change for different significance levels? Online Appendix A.7 (and a previous version of our paper, Lee et al. 2020) provides answers to these questions.

**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 \(\rho\) and \(f_0 \neq 0\). Deferring details to online Appendix B, we now outline the construction of such a function \(c_\alpha(F)\), which—as is apparent from Tables 3A and 3B—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

\(^{36}\)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.
probability occurs when $\rho = \pm 1$. This leads to obtaining a function $c_\alpha(F)$ such that the null rejection probability under $\rho = \pm 1$ is exactly equal to $\alpha$,

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

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

The following fact is central to our construction of the $tF$ critical value function: when $\rho = 1$, $t_{AR}$ is a linear function of $f_0$, and therefore equation (2) reduces to

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

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 $f_0'''$.

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 $\Phi(\bar{f}(f_0) - f_0) - \Phi(\bar{f}(f_0) - f_0)$, where the intersections between the two curves are denoted by $\bar{f}(f_0)$ and $\bar{f}(f_0)$. 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 $[\bar{f}(f_0), \bar{f}(f_0)]$.

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, $\bar{f}(f_0)$ and $\bar{f}(f_0)$, where

---

38 In online 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.
40 The figure uses the transformation $\left(\frac{|t^2/1.96|}{1 + (t^2/1.96)^2}\right)$ 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 $f_0 = f_0'''$).
\[ \Phi(f_0) - f_0 = 1 - \alpha. \]

This definition can be expressed more formally as the function \( \tilde{c}_\alpha(\cdot) \) satisfying the following system of equations:

\[
\begin{align*}
\frac{\tilde{f}(f_0)^2(\tilde{f}(f_0) - f_0)^2}{f_0^2} - \tilde{c}_\alpha(\tilde{f}(f_0)^2) &= 0, \\
\Phi(\tilde{f}(f_0) - f_0) - \Phi(f(f_0) - f_0) &= 1 - \alpha,
\end{align*}
\]

for a set of small values of \( f_0 \).

Whether or not there exists any continuous and decreasing function \( \tilde{c}_\alpha(f^2) \) satisfying this system of equations is not obvious and is technically challenging to prove. We defer those details to online 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 online Appendix B, as \( f^2 \downarrow q_{1-\alpha} \), the function \( \tilde{c}_\alpha(f^2) \) behaves as

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

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

\[ \text{Figure 7. Construction of the } t \text{ Critical Value Function} \]

\[ \text{Note: To aid visualization of values, vertical axis uses the transformation } \frac{(t^2/1.96^2)}{1 + (t^2/1.96^2)}. \]
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 \( \alpha = 0.05 \). We start with a set of points \( \left( f, \tilde{c}_{0.05}(f^2) \right) \), defined over the small interval \( f \in [-1.96 - \epsilon, -1.96] \) for \( \epsilon > 0 \). 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 \( f \), allowing \( f \) to be relabeled \( f(0) \). Then, the second equation in (6) can be used to solve for \( \tilde{f} \). Finally, the first equation in (6) can be used to solve for \( \tilde{c}_{0.05}(\tilde{f}(0))^2 \). This mapping produces a segment of the function defined on an interval that is longer than \( [-1.96 - \epsilon, -1.96] \). 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.

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 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, possesses 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 online Appendix B.1.

The second, and more straightforward, step in constructing the \( tF \) critical value function is to determine where, along function \( \tilde{c}_{0}(f^2) \), 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 \( \alpha \), for all values of \( \rho \) and \( f_0 \). In practice, we use numerical integration.

---

43 The W functions in Figure 7 do not depend on \( \alpha \), but the critical value curves do. The approach we outline here for \( \alpha = 0.05 \) can be applied more generally for other values of \( \alpha \). In online Appendix B, we give details regarding other values of \( \alpha \) 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.
of the expression in (4) to compute these rejection probabilities, as illustrated in Figure 2, for a grid of values for $f_0$ and $\rho$ to verify size control. 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 $A'B'C'$ 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 to the smallest possible value of that construction process, $2.73^2$. Online 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 online 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.

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 IIC, which showed that neither $AR$ nor $tF$ seemed to dominate across all values of $\Delta(\beta_0)$ or differing combinations of $\rho$ 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 $\beta_0$, where each value of $\beta_0$ in the parameter space is given equal weight. Power curves are conceived as rejection rates while keeping $\beta_0$ fixed while varying $\beta$, but our curves, since they are functions of $\Delta(\beta_0) = \frac{\sqrt{V(Z_v)}}{\sqrt{V(Z_u)}}(\beta - \beta_0)$, could equivalently be viewed as graphing power fixing $\beta$, while varying $\beta_0$. 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

---

45 Specifically, we use two grids. The first grid consists of values of $\rho$ 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 $\rho$ 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.
us to interpret the conditional expected length as the average type II error—averaged across all false hypotheses \( \beta_0 \)—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 \( \rho, f_0 \neq 0 \),

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

We next provide some intuition for this result. We show in online 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.,

\[
(8) \quad L_{AR} = \frac{\sqrt{F} \sqrt{F - q_{1-\alpha}(1 - \bar{\rho}^2)}}{F - q_{1-\alpha}} L_{IV},
\]

and

\[
L_{tF} = \sqrt{\frac{c_\alpha(F)}{q_{1-\alpha}}} L_{IV},
\]

where \( \bar{\rho}^2 = \frac{(-t_{AR}(\beta) + \rho f)^2}{(f^2 - 2\rho t_{AR}(\beta) f + t^2_{AR}(\beta))}. \)

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 online Appendix B.2 that

\[
\lim_{F \downarrow q_{1-\alpha}} c_\alpha(F)(F - q_{1-\alpha}) = q^3_{1-\alpha}.
\]

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. Online Appendix C provides the full proof.
In summary, these results show that the expected length of the \( t_F \) 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 \( t_F \) 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: \( t_F \) 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 nonconvex and unbounded confidence set.\(^{47}\) If, for example, in practice researchers effectively ignore the nonconvexity and simply use the whole real line as the confidence set, then the confidence sets for \( t_F \) 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 \( tF \).

**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 \( \pm 1.96 \) 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 \( \beta \), 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 mis-application 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

---

\(^{46}\) 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 nonconvex confidence set is reported other than Cruz and Moreira (2005).
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 reassess 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


This article has been cited by:

1. Rashmi Barua, Prarthna Agarwal Goel, Renuka Sane. 2023. Son preference and crime in India. *Review of Economics of the Household* 13. . [Crossref]


5. Joshua Angrist, Michal Kolesár. 2023. One instrument to rule them all: The bias and coverage of just-ID IV. *Journal of Econometrics* 20. . [Crossref]


7. Jaya Jumrani. 2023. How responsive are nutrients in India? Some recent evidence. *Food Policy* 114, 102379. [Crossref]

8. Rosanna Smart, Jacqueline Doremus. 2023. The kids aren’t alright: The effects of medical marijuana market size on adolescents. *Journal of Health Economics* 87, 102700. [Crossref]


10. Vitor Miguel Ribeiro. 2023. Green bond market boom: did environmental, social and governance criteria play a role in reducing health-related uncertainty?. *Green Finance* 5:1, 18-67. [Crossref]

11. David McKenzie, Dean Yang. *Field and Natural Experiments in Migration* 32, . [Crossref]


17. Natee Amornsiriapanitch. 2020. Why are Residential Property Tax Rates Regressive?. *SSRN Electronic Journal* 3. . [Crossref]