NBER WORKING PAPER SERIES

WHAT TO DO WHEN YOU CAN'T USE '1.96' CONFIDENCE INTERVALS FOR IV

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

Working Paper 31893
http://www.nber.org/papers/w31893

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2023

We continue to benefit from conversations we had with, and the work of, Charlie Fefferman. We are grateful to Jonathan Roth for comments, and also thank Bo Honoré, Michal Kolesár, Ulrich Müller, Mikkel Plagborg-Møller, Mark Watson, and participants of the Princeton Econometrics Workshop for their comments and suggestions. We are grateful to Ethan Bergmann, Santiago Deambrosi, Colin Dunkley, Monica Essig Aberg, Bernardo Esteves, Kyle Hancock, and especially Cate Brock for outstanding research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

What to do when you can't use '1.96' Confidence Intervals for IV
David S. Lee, Justin McCrary, Marcelo J. Moreira, Jack R. Porter, and Luther Yap
NBER Working Paper No. 31893
November 2023
JEL No. C01,C26,J0

ABSTRACT

To address the well-established large-sample invalidity of the +/-1.96 critical values for the t-ratio in the single variable just-identified IV model, applied research typically qualifies the inference based on the first-stage-F (Staiger and Stock (1997) and Stock and Yogo (2005)). We fully extend this F-based approach to its logical conclusion by presenting new critical values for the t-ratio to additionally accommodate values of F that do not meet existing thresholds needed for validity. These new t-ratio critical values simultaneously fix the main problem of over-rejection (invalidity) and the under-appreciated possibility of under-rejection (conservativeness) that can occur when relying solely on the usual 1.96 critical value. We show that the corresponding new confidence intervals are generally expected to be substantially shorter than competing "robust to weak instrument" intervals, including those from the recommended benchmark of Anderson and Rubin (1949) (AR). In a sample of 89 specifications from 10 recent empirical studies drawn from five general interest journals, the new "VtF" intervals are shorter than AR intervals 100 percent of the time, and even more likely to produce statistically significant results than the usual +/-1.96 procedure.

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
FGV EPGE
moreira.marcelo.j@gmail.com

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

Luther Yap
Princeton University
lyap@princeton.edu

Supplementary Material, including Appendix, and STATA code is available at
http://www.princeton.edu/~davidlee/wp/SupplementVtF.html

I Introduction

Consider the single-regressor, single excluded instrumental variable (IV) (just-identified) model, with outcome Y, regressor X and instrument Z,

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

where X can always be decomposed as $X = \pi Z + v$, and $\pi \equiv \frac{C(X,Z)}{V(Z)}$ is the population first-stage coefficient and v is the population least squares residual. This model[1] has seen use across a wide range of empirical disciplines.[2]

It has long been recognized that without any further assumptions, conventional t-ratio based inference – using the standard 2SLS point estimator and standard errors and the ± 1.96 critical values – produces invalid inference for the parameter β in model (1).[3] Put simply, even in large repeated samples, the t-ratio approximately follows a non-normal distribution (with known functional form), with departures from normality especially stark when the true (and unknown) value of $C(Z,X)$ is small. Due to the the weak-instrument asymptotic approximation of Staiger and Stock (1997), it is possible to quantify how anti-conservative the usual ± 1.96 procedure would be under different specific scenarios.

This inferential problem is the starting point for a great deal of work in the econometrics literature and is, at least implicitly, acknowledged in applied research: the current "industry standard" – highlighted in popular econometrics textbooks (e.g., Angrist and Pischke (2009), Stock and Watson (2019), Wooldridge (2019), Hansen (2022)), and provided by default in popular software packages (e.g., see Baum, Schaffer and Stillman (2003))– is to additionally report the observed first-stage F-statistic as a diagnostic guide to whether the usual t-ratio inference is problematic. The reporting of the first-stage F-statistic has intuitive appeal to researchers because observing a large F is more likely if the instrument is truly strong. As an example of its proper application, the F-threshold of 16.38 from Stock and Yogo (2005) for the just-identified case is derived so that the following is

[1] Covariates, including a constant, are easily accommodated by viewing X, Y, Z as residuals from regressing the original variables on the covariates and a constant. More generally, it is straightforward to accommodate covariates throughout the findings of this paper.

[2] The estimand β can be interpreted as the local average treatment effect (LATE) (Imbens and Angrist (1994)). This IV model is also employed to implement simple versions of fuzzy regression discontinuity or kink designs (Hahn, Todd and der Klaauw (2001), Lee and Lemieux (2010), and Card et al. (2015)).

[3] See, for example, Bound, Jaeger and Baker (1995), Dufour (1997), Nelson and Startz (1990), and Staiger and Stock (1997). For a recent survey of the econometric literature on weak instruments, see Andrews, Stock and Sun (2019).

true under the null hypothesis:

$$\Pr[\{F > 16.38\} \cap \{|t| > 1.96\}] \leq 0.15 \tag{2}$$

That is, the procedure of rejecting the null only when *both* the first stage F is larger than 16.38 and the t-ratio is larger than 1.96 in absolute value can be interpreted as a test at the 15 percent level of significance.[4]

Equivalently, this means that the use of the interval $\hat{\beta} \pm 1.96 \cdot \hat{se}(\hat{\beta})$ when $F > 16.38$ (and setting the confidence set to include all values on the real line when $F < 16.38$) is to be viewed as an 85 percent confidence interval.[5] The use of the F-statistic in this way, therefore, dramatically improves confidence levels since the use of $\hat{\beta} \pm 1.96 \cdot \hat{se}(\hat{\beta})$ (ignoring F) has 0 percent confidence level, as established by Dufour (1997). The analogous first-stage F threshold to obtain 5/95 percent significance/confidence is a somewhat higher value of 104.67 (Lee et al. (2022)). Lee et al. (2022) (henceforth, LMMP (2022)) further refine the method of Stock and Yogo (2005) (henceforth, SY (2005)) so that even if $F < 104.67$, and indeed even if F is lower than 16.38 (or the commonly-cited value of 10) – as long as F is larger than 1.96^2 – valid and bounded confidence intervals are still possible by changing the 1.96 scaling of standard errors in the usual confidence intervals to a scaling that is a smooth function of the first-stage F, which is presented in LMMP (2022).[6]

In this paper, we extend the F-based approach of SY (2005) and its tF refinement (LMMP (2022)) to its logical conclusion, resulting in a new t-ratio-based procedure, which we call VtF (because tF of LMMP (2022) is a special case of VtF, with a "V" to indicate it is using variance information, as explained below). VtF not only eliminates the possibility of over-rejection – the first order problem with IV inference; it also eliminates an underappreciated property of the usual ± 1.96 critical values – they also have the potential to lead to excessively *conservative* inferences.

[4]It is important to note here that the unconditional probability in (2) is distinct from the conditional probability $\Pr[\{|t| > 1.96\} | F > 16.38]$. The unconditional rejection probability in (2) is the standard focus of the weak-IV literature, and in particular the focus of Stock and Yogo (2005) for calculating the relevant critical values for F, as well as the focus of Staiger and Stock (1997) in their discussion of how the F-statistic can be incorporated into inference. Accordingly the scope of this paper follows this standard. By contrast, the conditional probability describes the rejection probability under a process in which decisions about acceptance or rejection are made only for those realizations where F is at least 16.38 for example (and otherwise the data is discarded and hence no inference is made). This "screening" phenomenon is discussed in Andrews, Stock and Sun (2019) and more recently explored in Angrist and Kolesár (2023), who find that if one commits to discarding the data based on the sign of the first-stage coefficient (i.e., focus on the conditional probability $\Pr[\{|t| > 1.96\} | \hat{\pi} > 0]$), the possible risk of over-rejection, compared to the original (untruncated) inferential problem, does not change very much. Separately, Angrist and Kolesár (2023) also contains an analysis of the standard unconditional problem, which we discuss in greater detail below.

[5]The significance/confidence level given here is based on a Bonferroni bounding approach discussed in Staiger and Stock (1997). A precise calculation (using the formulas in Lee et al. (2020)) shows that the significance/confidence associated with the 16.38 threshold is 9.33/90.67 percent.

[6]See http://www.princeton.edu/~davidlee/wp/SupplementarytF.html for STATA code that provides tF critical values from LMMP (2022). For STATA code that provides VtF critical values, see http://www.princeton.edu/~davidlee/wp/SupplementVtF.html

The implementation of VtF is simple: use the usual 2SLS point estimate and standard error, and hence the usual t-ratio, and instead of ± 1.96 use the new critical values we present. These critical values depend on the first-stage F-statistic – just as in SY (2005) and LMMP (2022). But in addition, the critical values depend on another easily computable quantity, the sample correlation of the main equation and first-stage residuals, while imposing the null $\beta = \beta_0$, denoted $\hat{\rho}(\beta_0)$. We construct the new critical value function $\sqrt{c\left(\hat{\rho}(\beta_0), \hat{F}\right)}$ to have the desired property that when the null is true,

$$\lim_{N\to\infty} \Pr\left[|\hat{t}| > \sqrt{c\left(\hat{\rho}(\beta_0), \hat{F}\right)}\right] = 0.05.$$[7]

This statement is true whether the instrument is strong or arbitrarily weak. All test procedures necessarily lead to a corresponding confidence set procedure (e.g. the usual t-ratio test $\left|\frac{\hat{\beta}-\beta_0}{\hat{se}(\hat{\beta})}\right| > 1.96$ leads to the usual confidence intervals $\hat{\beta} \pm 1.96 \cdot \hat{se}(\hat{\beta})$). In this case, we show that the corresponding confidence set will be contained by a simple-to-compute interval of the form

$$\left[\hat{\beta} - k^{-}\left(\hat{r}, \hat{F}\right) \cdot \hat{se}(\hat{\beta}), \hat{\beta} + k^{+}\left(\hat{r}, \hat{F}\right) \cdot \hat{se}(\hat{\beta})\right] \tag{3}$$

where $\hat{r}$ is the sample correlation between the main equation and first-stage residuals (from using $\hat{\beta}$), and $k^{-}(\cdot,\cdot)$ and $k^{+}(\cdot,\cdot)$ are functions that we derive and present below.[8]

Our development of the VtF procedure and investigation of its properties leads to the following contributions, which should be of interest to both applied researchers and econometricians who are invested in the just-identified instrumental variable model.

First, like LMMP (2022) and SY (2005) before it, VtF critical values cater to practitioners' apparent preference for the familiar ± 1.96 confidence interval construction. But it extends the justification of the use of the usual ± 1.96 intervals, and reduces the frequency with which further adjustment is needed to ensure the intended 95% confidence level. Instead of requiring $\hat{F} > 104.67$ (LMMP (2022)), an expanded and simple rule of thumb $\hat{F} > 10 + 100 \cdot \hat{r}$ is sufficient for relying on the ± 1.96 intervals for 95% confidence (LMMP (2022)). This is made possible because when this

[7] Throughout the paper, we focus on the case of 5% significance or 95% confidence levels, but we also provide the critical values and confidence interval adjustment factors for the 1/99 percent significance/confidence levels. Note also that our approach accommodates the commonly-employed departures from i.i.d. errors, as long as the appropriate heteroskedasticity-consistent, clustered, or time series variance estimators are used.

[8] More precisely, $\hat{r} = \frac{\hat{\rho}_{RF} - \hat{\beta}\sqrt{\frac{\hat{\sigma}_{22}}{\hat{\sigma}_{11}}}}{\sqrt{1 - 2\hat{\rho}_{RF}\hat{\beta}\sqrt{\frac{\hat{\sigma}_{22}}{\hat{\sigma}_{11}}} + \left(\hat{\beta}\sqrt{\frac{\hat{\sigma}_{22}}{\hat{\sigma}_{11}}}\right)^2}}$, where $\hat{\beta}$ is the 2SLS estimator, and $\hat{\sigma}_{11}, \hat{\sigma}_{22}$, and $\hat{\rho}_{RF}$ constitute the consistent, robust (to e.g. heteroskedasticity, clustered errors) variances and correlation of the reduced-form and first-stage coefficients. This is essentially equation 6, after replacing β with $\hat{\beta}$ and replacing σ_{11}, σ_{22}, and ρ_{RF} with $\hat{\sigma}_{11}, \hat{\sigma}_{22}$, and $\hat{\rho}_{RF}$. It can be shown that when these reduced-form variances are the homoskedasticity-only version, $\hat{r}$ simplifies to the sample correlation between $\hat{u}$ and $\hat{v}$. Under the heteroskedasticity-consistent variance estimator, $\hat{r}$ simplifies to the sample correlation between $Z\hat{u}$ and $Z\hat{v}$. Under the clustered-error-consistent variance estimator, $\hat{r}$ simplifies to the sample correlation across cluster groups of the summation of $Z\hat{u}$, $Z\hat{v}$ within groups.

condition is satisfied, VtF intervals in (3) are *entirely contained within* the usual ± 1.96 intervals – notably something that the AR statistic of Anderson and Rubin (1949) never achieves. And even when the rule of thumb is not satisfied – $\hat{F}$ could be as low as $1.96^2 = 3.84$ – VtF produces usable confidence intervals while maintaining 95 percent confidence.

Second, we compare the performance of VtF to other valid 95% confidence procedures (AR and tF) as applied to 89 just-identified IV regressions from a sample of empirical studies recently published in five general interest economics journals (*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and the *Review of Economic Studies*). The results are clear-cut. In all 89 specifications, VtF yields shorter confidence intervals than both AR and tF intervals, often by a substantial margin. Furthermore, we document that in this sample, VtF is the most successful in producing statistically significant results.

Third, even under the restricted conditions that justify the usual ± 1.96 procedure, VtF delivers more precise inferences. Specifically, these conditions (see Angrist and Kolesár (2023)) are 1) β_0 lies in a particular, bounded "valid zone" (which is different across datasets), or 2) *a priori* knowledge is used to bound β and those bounds are contained within the "valid zone". We show that under either of these conditions the VtF critical values and inflation factors, compared to the usual ± 1.96 procedure, will more frequently detect departures from the null, and will produce shorter confidence intervals. Thus, VtF not only allows valid inferences when the usual procedure cannot, but even in the situations in which the usual procedure is valid, VtF leads to more precise inferences.

Fourth, to assess whether or not the above results are driven by the idiosyncracies of our sample, we introduce a new way to exhaustively and transparently characterize relative confidence interval performance (and implicitly, relative power) across all procedures. Inference methods are typically analyzed using power curves, which yield "on average" comparisons of the methods for particular data generating processes. We provide a much finer comparison that involves no commitment to a data generating model, averages, or any kind of approximation, and instead we compare confidence sets (up to a shared location and scale normalization) for *any* data set. In particular, we show that the relative confidence interval lengths (and relative positions) of all the methods for any dataset depend on only two statistics, $\hat{F}$ and $|\hat{r}|$. This dimension reduction allows us to present a comparison of confidence interval lengths using a two-dimensional heat map. From this heat map, it is easy to see why VtF may very frequently outperform AR in practice, as it does in our sample: the subset of $(\hat{F}, \hat{r})$ values such that AR intervals are shorter than VtF is relatively small. The heatmap also helps us to determine, for example, that there exists no dataset that will produce a VtF 95% confidence interval that is more than 8.8 percent longer than the 95% AR interval. By contrast, there is no bound to how much longer AR intervals can be, relative to VtF intervals. As expected, VtF confidence intervals are always shorter than that of tF from LMMP (2022).

And for relatively small F-statistics, tF intervals are substantially longer than VtF intervals and roughly speaking, AR intervals are longer than VtF intervals by the same degree that tF intervals (which do not use $\hat{r}$) are longer than AR intervals (which do implicitly use $\hat{r}$). An immediate implication of these findings for applied work is that IV inferences that only rely on the F-statistic can be made substantially more precise through reporting of one additional statistic, $\hat{r}$, something that is easy to do – and recoverable from reporting the point estimates and standard errors from three regressions (2SLS, reduced-form, first-stage)[9] – and yet only done in a minority of recent published studies that use IV.

Fifth, since our findings run counter to the conventional econometric "folk wisdom" that applied research should abandon t-ratio based inference and instead employ Anderson-Rubin (AR), we also present a traditional power comparison of methods to reconcile the VtF performance shown here with the previous econometrics literature.[10] After noting that VtF is outside the class of tests within which AR is uniformly most powerful, we show that, generally, across various specifications the relative power advantages of VtF and AR appear roughly balanced. This appearance of balance diminishes as the correlation between the main equation and first stage error decreases, and VtF's advantage over AR becomes more apparent. When the correlation is exactly zero, VtF is more powerful than AR over all departures of β from the null β_0. There appears to be no design for which the reverse is true. More importantly, we observe that the traditional power analysis averages power when realizations of the data lead to unbounded confidence sets with power when realizations of the data lead to bounded confidence sets. When we decompose the power curves to the separate cases of unbounded and bounded confidence sets, we find that AR is relatively more powerful in the event of obtaining statistically insignificant first-stage F values (when the confidence sets are unbounded). On the other hand, VtF generally possesses a decided power advantage in the event of statistically significant first-stage F values (when confidence sets are bounded). This enhanced power curve analysis isolates the sources of our earlier findings on VtF's superior confidence set length.

The paper is organized as follows. Section II provides a high-level summary of existing procedures that accommodate small values of F, and also illustrates the relative performance of VtF

[9] Footnote 8 shows how $\hat{r}$ is computed using $\hat{\beta}$ and $\hat{\rho}_{RF}$ and $\frac{\hat{\sigma}_{22}}{\hat{\sigma}_{11}}$. It can also be shown that, given $\hat{r}$, the consistent estimators of $\hat{\rho}_{RF}$ and $\frac{\hat{\sigma}_{22}}{\hat{\sigma}_{11}}$ can be recovered through $\hat{\rho}_{RF} = \frac{\hat{\beta}|\hat{\pi}|\frac{\sqrt{\hat{\sigma}_{22}}}{\sqrt{N}\hat{se}(\hat{\beta})}+\hat{r}}{\sqrt{\frac{\hat{\sigma}_{22}}{\left(\sqrt{N}\hat{se}(\hat{\beta})\right)^2}\hat{\pi}^2\hat{\beta}^2+2\hat{r}\hat{\beta}|\hat{\pi}|\frac{\sqrt{\hat{\sigma}_{22}}}{\sqrt{N}\hat{se}(\hat{\beta})}+1}}$ and $\frac{\hat{\sigma}_{22}}{\hat{\sigma}_{11}} = \frac{\frac{\hat{\sigma}_{22}}{\left(\sqrt{N}\hat{se}(\hat{\beta})\right)^2}\hat{\pi}^2}{\frac{\hat{\sigma}_{22}}{\left(\sqrt{N}\hat{se}(\hat{\beta})\right)^2}\hat{\pi}^2\hat{\beta}^2+2\hat{r}|\hat{\pi}|\frac{\sqrt{\hat{\sigma}_{22}}}{\sqrt{N}\hat{se}(\hat{\beta})}\hat{\beta}+1}$.

[10] See Andrews, Stock and Sun (2019) for a recent comprehensive survey of the econometrics literature on this topic.

using our sample of empirical studies. Section III defines and describes the VtF procedure for hypothesis testing and confidence intervals and presents our main theoretical results. Section IV presents a power analysis to help explain why VtF outperforms AR and also how our findings can be reconciled with results from the previous econometric literature. Section V concludes.

II VtF: Motivation and Summary of Performance in Practice

In this section, we provide a high-level motivation for the new critical values that we compute (henceforth, VtF critical values), and a summary of their impact and performance in practice, comparing VtF to alternative available approaches. We do so by applying all of the inference procedures to a sample of IV specifications drawn from recently published articles in economics. We go beyond this applied perspective and provide a more formal discussion in Sections III and IV.

We now describe our sample: it contains reported statistics from every single-variable just-identified IV specification appearing in articles published in the *American Economic Review*, *Econometrica,* the *Journal of Political Economy,* the *Quarterly Journal of Economics*, and the *Review of Economic Studies* in the year 2021. A "specification" refers to a unique combination of dependent variable, endogenous regressor of interest, set of covariates, and single excluded instrument. There are a total of 418 published articles (excluding comments and replies) in these five journals during 2021. 89 of them are classified as having used "instrumental variables". Of these, 69 were identified as containing at least one just-identified specification. And from this group, fourteen studies were identified as reporting the equivalent of the following statistics from three regressions:

1. The 2SLS point estimate $\hat{\beta}$ and standard error $\hat{se}(\hat{\beta})$
2. The first-stage estimate $\hat{\pi}$ and standard error $\hat{se}(\hat{\pi})$
3. The reduced-form point estimate $\widehat{\pi\beta}$ and standard error $\hat{se}(\widehat{\pi\beta})$

With these statistics in hand, we are able to compute the VtF critical values and implement alternative inference methods for comparison without access to the complete data sets. It is most common for researchers to report statistics from regressions 1) and 2), and less common for researchers to report statistics from all three regressions, even though they are easily computable.[11] In summarizing these data, throughout our analysis, we describe the data at the level of the specification, using weighted proportions, means, and percentiles, where the weights for each specifi-

[11] In the sample of LMMP (2022), about one-third of the specifications report statistics from regressions 1) and 2), and one-quarter report statistics from all three regressions.

cation is the reciprocal of the number of specifications within a particular study, so as to give each study equal weight.

In the full sample of 109 specifications from these fourteen studies, the (weighted) median first-stage F statistic for the 109 specifications is about 47, while the 25th and 75th percentiles are about 15 and 190, respectively.[12] We focus on two main subsamples from these 109 specifications in this paper. First, in the next two subsections, for discussing alternative options for when the F-statistic is too small for single-threshold rules (e.g. SY (2005)) to be informative, we focus on the 62 specifications (drawn from eight distinct studies) for which the F-statistic is less than 18. Later, in Section III, when we comprehensively document confidence length comparisons between various procedures, we focus on the the 89 specifications (drawn from 10 distinct studies) for which the F-statistic is between 3.84 and 104.67, since the practitioner using either tF or VtF can revert to the usual ± 1.96 intervals when F exceeds 104.67.

II.A Current Options for *IV* inference when F is "small"

The most commonly employed approach to IV inference in applied research is to consider whether the first-stage F-statistic is sufficiently high, say, greater than some threshold $\bar{F}$. Although SY (2005) introduce thresholds $\bar{F}$ for the first-stage F for the purpose of formally testing the null hypothesis that "the instrument is weak", the thresholds can also be used, in conjunction with the argument put forth in Staiger and Stock (1997), to modify the IV t-ratio procedure so that it delivers inference at a known level of significance/confidence. Specifically, the proposal there, applied to the thresholds in SY (2005) leads the researcher to the current *de facto* industry standard: use the usual confidence interval $\hat{\beta} \pm 1.96 \cdot \hat{se}(\hat{\beta})$ if $\hat{F} > \bar{F}$, and accept all possible values of β if $\hat{F} < \bar{F}$ (i.e., the confidence set is the whole real line if $\hat{F} < \bar{F}$). The choice of the threshold $\bar{F}$ determines the confidence level of the interval. If $\bar{F} = 10$ the confidence interval has 88.6 percent confidence level, while for $\bar{F} = 16.38$, the confidence level is 90.6. As shown in LMMP (2022), a 95 percent confidence level requires $\bar{F} = 104.67$.

The main limitation to the single-threshold approach is that "relatively small" F-statistics are a common occurrence in practice. For example, in our sample, 35 percent of the F-statistics are below 18. Consider a researcher who observes a t-statistic for the first stage around 3 or 4 (i.e., a first-stage F around the range of 9 to 16); this would typically be viewed as strong evidence of the existence of a first stage relationship. Despite this, the use of $\bar{F} = 16.38$ or $\bar{F} = 104.67$ would lead them to accept that they cannot learn anything from the instrumental variable strategy. It would be desirable to be able to make informative inferences in these circumstances:

[12] We note that the LMMP (2022) sample is not restricted to those reporting statistics from all three regressions. The 25th, 50th, 75th percentiles of the first-stage F from that sample are 14.23, 45.84, and 225, respectively.

when $\hat{F}$ indicates a statistically significant first-stage, but is not large enough to achieve intended significance/confidence levels.

Table 1 summarizes three existing and very different inference approaches that accommodate small F-statistics. We describe each below, and also draw attention to each of their key limitations. The first option, summarized in the first row of Table 1 is the procedure of Anderson and Rubin (1949), which rejects the null hypothesis if the AR statistic is greater than 1.96^2. In the just-identified model, the AR statistic is equivalent to the score/Lagrange multiplier and the likelihood ratio statistic for the hypothesis $\beta = \beta_0$. When $\hat{F} > 1.96^2$, the AR confidence set is an interval and can be written as

$$\left[\hat{\beta} - k_{AR}^{-}\left(\hat{r},\hat{F}\right)\cdot \hat{se}(\hat{\beta}), \hat{\beta} + k_{AR}^{+}\left(\hat{r},\hat{F}\right)\cdot \hat{se}(\hat{\beta})\right]$$

where $k_{AR}^{+}(\cdot,\cdot)$ and $k_{AR}^{-}(\cdot,\cdot)$ are known functions of the observed statistics $\hat{r}$, $\hat{F}$.[13] This approach is rarely used in practice (LMMP (2022)), even though it is fully robust to any degree of instrument strength and has some optimality properties, a point that we discuss and qualify in greater detail in Section IV. One important practical difference between AR and the "F-based" approach of Stock/Yogo, is that the former does not collapse to the familiar ± 1.96 confidence intervals under any circumstances, even when the F-statistic is very large.[14]

A second approach that can also accommodate $\hat{F}$ values as low as 1.96^2, is the tF procedure (LMMP (2022)), summarized in the second row in Table 1. Essentially, tF is a refinement of the original Stock-Yogo F-based approach. Any given hypothesis β_0 is rejected at the 5 percent level if and only if $|\hat{t}| > \sqrt{c_{tF}\left(\hat{F};0.05\right)}$ and the 95 percent confidence interval is given by $\hat{\beta} \pm \sqrt{c_{tF}\left(\hat{F};0.05\right)}\cdot \hat{se}\left(\hat{\beta}\right)$, where $\sqrt{c_{tF}\left(\hat{F};0.05\right)}$ is a continuous decreasing function of $\hat{F}$, with values reported in LMMP (2022) (and available as STATA code). As with AR, confidence sets are unbounded when $\hat{F} \leq 1.96^2$ (when the first-stage coefficient is statistically insignificant), but are otherwise bounded intervals. Unlike AR, tF can be viewed as an extension of the familiar ± 1.96 procedure in the spirit of SY (2005): as long as $\hat{F}$ exceeds 104.67, use ± 1.96 critical values, and otherwise use as critical values $\pm\sqrt{c_{tF}\left(\hat{F};0.05\right)}$. On the other hand, tF shares with the SY (2005) method an inherent conservatism: validity necessitates that it remains valid even under the "worst case"/extreme scenario that the correlation $\rho \equiv Corr\left(Zv,Zu\right)$ has a value of ± 1. To the extent that the true ρ is relatively small, the test procedure will be conservative (i.e., reject strictly less than 5 percent of the time, under the null) or equivalently, produce excessively long confidence intervals

[13] See footnote 8.

[14] As noted in Angrist and Kolesár (2023), AR intervals are always longer than the usual ± 1.96 intervals. Although it would be tempting to adopt the procedure of the usual ± 1.96 intervals when $\hat{F}$ is large enough, and otherwise use AR, LMMP (2022) show that this "hybrid" test also does not achieve the intended 5 percent; any such hybrid will have a higher level of distortion than the usual Stock-Yogo approach given the same threshold $\bar{F}$.

Table 1: Alternatives for Valid IV Inference accommodating small F statistics

	Testing $H_0 : \beta = \beta_0$		Confidence Interval	
Method (Add. Stats Used)	Limitation on β_0	Test procedure: Reject iff	Additional Assumptions	Interval
AR ($\hat{se}(\hat{\pi\beta})$ or $\hat{r}$)	None	$AR > 1.96^2$	None	$\left[\hat{\beta} - k_{AR}^{-}\left(\hat{r},\hat{F}\right)\cdot \hat{se}(\hat{\beta}),\ \hat{\beta} + k_{AR}^{+}\left(\hat{r},\hat{F}\right)\cdot \hat{se}(\hat{\beta})\right]$
tF (None)	None	$\lvert\hat{t}\rvert > \sqrt{c_{tF}\left(\hat{F}\right)}$	None	$\hat{\beta} \pm \sqrt{c_{tF}\left(\hat{F}\right)}\cdot \hat{se}(\hat{\beta})$
"Limited" t ($\hat{se}(\hat{\pi\beta})$ or $\hat{r}$) (see Angrist and Kolesár (2023))	Test valid only for $\beta_0 \in \left[\hat{\rho}^{-1}(.565),\ \hat{\rho}^{-1}(-.565)\right]$	$\lvert\hat{t}\rvert > 1.96$	Need *a priori* bounds β_{lower} and β_{upper} with $\hat{\rho}^{-1}(.565) \leq \beta_{lower}, \beta_{upper} \leq \hat{\rho}^{-1}(-.565)$	$\left[\hat{\beta} - 1.96\hat{se}(\hat{\beta}),\ \hat{\beta} + 1.96\hat{se}(\hat{\beta})\right] \cap \left[\beta_{lower}, \beta_{upper}\right]$
VtF ($\hat{se}(\hat{\pi\beta})$ or $\hat{r}$)	None	$\lvert\hat{t}\rvert > \sqrt{c\left(\hat{\rho}\left(\beta_0\right),\hat{F}\right)}$	None	$\left[\hat{\beta} - k^{-}\left(\hat{r},\hat{F}\right)\cdot \hat{se}(\hat{\beta}),\ \hat{\beta} + k^{+}\left(\hat{r},\hat{F}\right)\cdot \hat{se}(\hat{\beta})\right]$

Note: All methods require at least $\hat{\beta}$, $\hat{se}(\hat{\beta})$, $\hat{\pi}$, and $\hat{se}(\hat{\pi})$, and additional statistics needed for each approach are indicated in the first column. Given $\hat{\beta}$, $\hat{se}(\hat{\beta})$, $\hat{\pi}$, and $\hat{se}(\hat{\pi})$, one can recover $\hat{r}$ from $\hat{se}(\hat{\pi\beta})$ and vice versa.

(i.e., cover the true parameter more than 95 percent of the time).[15]

Summarized in the third row of Table 1, a third, novel approach, recently introduced by Angrist and Kolesár (2023), finds a different way to adhere to the usual ± 1.96 critical values whenever possible. It involves abandoning the use of the F-statistic altogether and instead focuses on the fact that, for any given problem, there is a specific subset values for β such that the usual $|\hat{t}| > 1.96$ procedure does not over-reject.[16]

More specifically, they leverage a subtle, and under-appreciated fact: there exists a one-to-one relation between ρ and β that is fully determined by the variance-covariance matrix of the first-stage and reduced form estimators. Denoting this one-to-one mapping as $\rho(\beta)$, this means that any hypothesized value β_0 necessarily implies a hypothesized correlation $\rho(\beta_0)$. This function, and its inverse, are consistently estimable and denoted as $\hat{\rho}(\cdot)$ and $\hat{\rho}^{-1}(\cdot)$.[17]

Angrist and Kolesár (2023) use the fact that the test $|\hat{t}| > 1.96$ properly rejects no more than 5 percent of the time as long as $|\rho| \leq .565$.[18] Thus, as long as the hypothesis of interest β_0 lies in the

[15] That tF is conservative relative to AR is pointed out in Angrist and Kolesár (2023) and Keane and Neal (2023).

[16] In a separate point about how to consider inference *conditional* on discarding the data when the F-statistic is too small (i.e. truncated/screened inference), Angrist and Kolesár (2023) discuss the use of the sign of the first stage coefficient in this context. See footnote 4.

[17] The functional form of $\rho(\cdot)$ is given below in equation (6), and the functional form of $\hat{\rho}(\cdot)$ follows by replacing the covariance parameters in (6) with corresponding sample estimates.

[18] Both LMMP (2022) and Angrist and Kolesár (2023) use the same formulas from SY (2005) and independently confirmed the .565 number for the 5 percent level of significance.

interval $\left[\hat{\rho}^{-1}(.565), \hat{\rho}^{-1}(-.565)\right]$, then no modification to the usual ± 1.96 procedure is needed for tests at the 5 percent level of significance.

There are two limitations to this approach to inference. First, only certain values of β_0 can be tested. That is, the hypothesis of interest may not lie in the "valid zone" of $[\hat{\rho}^{-1}(.565), \hat{\rho}^{-1}(-.565)]$. As an example, the hypothesis $\beta_0 = 0$ is quite commonly a hypothesis of interest; in our subsample of 62 specifications, for 29 percent of them, zero lies outside the interval $[\hat{\rho}^{-1}(.565), \hat{\rho}^{-1}(-.565)]$, which thus leads to an invalid test. More frequently, for 52 percent of the specifications in our sample, the null hypothesis of 0 lies outside the $[\hat{\rho}^{-1}(.435), \hat{\rho}^{-1}(-.435)]$, which invalidates the more stringent test at the 1 percent level of significance using the usual ± 2.58 critical values.[19] Using a completely different sample of published studies (from the *American Economic Review*, LMMP (2022) report similar frequencies with which this approach would not be possible for testing the null that $\beta = 0$.[20]

The second limitation to this approach is that the corresponding confidence interval procedure would require imposing additional assumptions beyond the standard model (1). Specifically, it would be necessary to restrict the values of β to lie within an interval $\beta_{lower} \leq \beta \leq \beta_{upper}$, perhaps based on theory or findings from other empirical studies (as was done in Angrist and Kolesár (2023)). Next, to ensure validity, it would be necessary for those bounds to satisfy $\hat{\rho}^{-1}(.565) \leq \beta_{lower}$ and $\beta_{upper} \leq \hat{\rho}^{-1}(-.565)$. Under these conditions, the valid confidence interval is given by

$$\left[\hat{\beta} - 1.96 \cdot \hat{se}\left(\hat{\beta}\right), \hat{\beta} + 1.96 \cdot \hat{se}\left(\hat{\beta}\right)\right] \cap [\beta_{lower}, \beta_{upper}]$$

where there is a possibility that the usual data-based interval is truncated by the *a priori* bounds $[\beta_{lower}, \beta_{upper}]$. The practitioner might hope that these bounds would be wide enough so that the confidence intervals are determined by the data, and not by the *a priori* bounds. This does not appear to be true for many studies in our sample. Figure 1 displays the lower and upper endpoints of the usual ± 1.96 confidence intervals for all 62 specifications.[21] Most of them are truncated by the "valid zone", given by the range $[\hat{\rho}^{-1}(.565), \hat{\rho}^{-1}(-.565)]$.[22,23]

[19] LMMP (2022), using the same formulas as used by LMMP (2022) and Angrist and Kolesár (2023) to compute the .565 number, report .435 as the analogous number for the 1 percent level, noting that the range of ρ for which the usual procedure remains valid depends on the desired level of statistical significance.

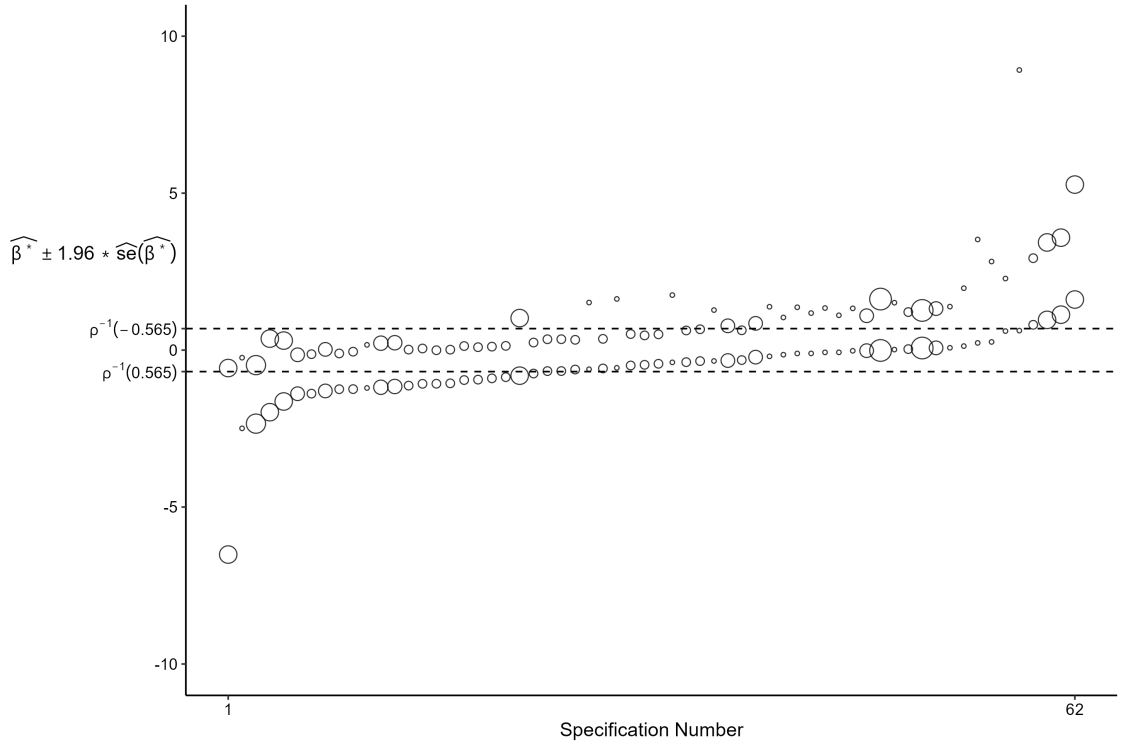
[20] Section A.8.2. of the Online Appendix reports that for 30 percent of the studies, the hypothesis $\beta = 0$ would lie outside the "valid zone" for the 5 percent test, with 42 percent lying outside the "valid zone" for the 1 percent test.

[21] We normalize the units of the confidence interval endpoints according to the linear transformation $\beta^* = \frac{\beta\sqrt{\frac{\hat{\sigma}_{22}}{\hat{\sigma}_{11}}} - \hat{\rho}_{RF}}{\sqrt{1-\hat{\rho}_{RF}^2}}$, which then leads to the simplification that $\rho = -\frac{\beta^*}{\sqrt{1+\beta^{*2}}}$ and $\beta^* = -\frac{\rho}{\sqrt{1-\rho^2}}$.

[22] 10 out of the 62 specifications (8.8 (weighted) percent) produce a ± 1.96 confidence interval that is not truncated by the "valid zone". Thus, we do find some examples of cases, like the three examples in Angrist and Kolesár (2023), where the confidence interval is not truncated by the "valid zone" for the 95% level. All of the usual 99% confidence intervals are truncated by the corresponding 99% "valid zone".

[23] A simple fix that would avoid the need to impose *a priori* bounds would be to switch to tF for the unavoidable

Figure 1: $\pm$ 1.96 Confidence Intervals vs *A Priori* Bounds



Note: All confidence bounds are normalized to units of β^*. Specifications are in order of the value of the lower bound of the usual $\pm$1.96 confidence interval. Circles indicate the endpoints of the confidence interval, and the size of the circles are proportional to the inverse of the number of specifications in each study. Horizontal dotted lines indicate the *a priori* bounds β_{lower} and β_{upper} required to justify the validity of the usual $\pm$1.96 confidence intervals.

The limitations of these three approaches motivate the development of VtF: we seek to 1) extend and enhance the original SY (2005) and LMMP (2022) approach of an F-dependent critical value function to accommodate both small F-statistics and the possibility of reverting to the usual $\pm$1.96 intervals when possible, 2) eliminate the inherent conservatism of these F-based approaches, which do not utilize all of the statistics of the model, while 3) accommodate hypothesis tests of any value of β, and avoid the need to specify *a priori* bounds on the parameter of interest.

In this paper, we derive a test, which we call VtF (summarized in the final row of Table 1): essentially, the procedure is the t-ratio inference based on corrected critical values. Specifically, we show that for any given hypothesis β_0 and associated $\hat{\rho}(\beta_0)$ there exists an F-dependent critical value function for the t-ratio that delivers rejection at the intended significance level for all possible values of the parameters of the model.[24] The test rejects if and only if $|\hat{t}| > \sqrt{c\left(\hat{\rho}\left(\beta_0\right), \hat{F}\right)}$, with values of the function $\sqrt{c\left(\hat{\rho}\left(\beta_0\right), \hat{F}\right)}$ provided in Appendix Table A3; the confidence set is

subset of β_0 values that would land outside the "valid zone". This approach to inversion for confidence intervals would also be a valid procedure. We thank Michal Kolesár for this observation; it captures the spirit and intuition of VtF.

[24] Section D.1 provides a precise description of the VtF critical value function uniqueness.

bounded when $\hat{F} > 1.96^2$ and is contained by the interval

$$\left[\hat{\beta} - k^{-}\left(\hat{r},\hat{F}\right)\cdot \hat{se}(\hat{\beta}), \hat{\beta} + k^{+}\left(\hat{r},\hat{F}\right)\cdot \hat{se}(\hat{\beta})\right]$$

where the functions $k^{-}(\cdot,\cdot)$ and $k^{+}(\cdot,\cdot)$ are provided in Appendix Table A5. Critical values and the confidence interval inflation factors for any values of $\hat{F}$, $\hat{\rho}(\beta_0)$, and $\hat{r}$ are available via STATA code provided at http://www.princeton.edu/~davidlee/wp/SupplementVtF.html.

We note here that the above confidence interval, like the AR interval, need not be symmetric around $\hat{\beta}$; as we discuss in more detail in Section III.C, it is easy to conservatively accommodate a preference for reporting a symmetric interval for both AR or VtF.

II.B Impact and Performance of VtF in Practice

The remaining sections of the paper provide a systematic treatment of the properties of VtF, but in this subsection we demonstrate a main finding of the paper by applying the procedure to our sample of IV specifications. As in our theoretical analysis below, we find in our sample of studies that VtF is more powerful than the above three existing valid alternatives. That is, VtF is more likely to result in statistically significant results, and its confidence intervals are considerably shorter as well.

Table 2 reports the frequency with which the null hypothesis that $\beta_0 = 0$ is rejected at the 5 percent level of significance using AR, tF, and VtF. In order to include the usual ± 1.96 procedure in the comparison, for this exercise, we further restrict the sample to those 51 specifications for which the usual test would be valid (i.e., the specifications for which $0 \in \left[\hat{\rho}^{-1}(.565), \hat{\rho}^{-1}(-.565)\right]$).

The first column focuses on the 38 specifications drawn across 7 studies for which the null hypothesis was rejected by the ± 1.96 rule. Among these, AR also rejected the null for almost all of them. tF, on the other hand, rejected about 59 percent of the time. VtF rejects the null 100 percent of the time for these studies, and the substantial difference between tF and VtF rejection rates clearly demonstrates the latter's power advantage in practice.

The second column focuses on the performance of the three alternatives among the 13 specifications (drawn from 4 distinct studies) for which the coefficient was statistically insignificant via the usual t procedure. Neither AR nor tF is able to reject the null in any of those cases. By contrast, among these cases VtF rejects the null for about 22 percent of these cases. Overall, in this restricted sample of specifications for which the usual t procedure, AR, and tF are all valid, VtF emerges as the most successful in yielding statistically significant results. The final column in Table 2 also shows that in the subset of 62 specifications (with F-statistics less than 18) from Figure 1, about 18 percent of the time the VtF intervals are shorter than the usual ± 1.96 intervals.

In our sample of specifications, VtF confidence intervals outperform that of tF and AR, the

Table 2: Frequency of Statistical Significance and Confidence Interval Length: t, AR, tF, VtF

Procedure	Reject (Proportion)		CI Is Shorter (Proportion)
	$\|t\| > 1.96$	$\|t\| < 1.96$	
AR	0.984	0.000	0.000
tF	0.588	0.000	0.000
VtF	1.000	0.223	0.179
N	38	13	62
(# of Studies)	7	4	8
% (Weighted) of Studies	78.7	21.3	

Note: All proportions weighted by the inverse of the number of specifications in each study. (# of Studies) indicates how many distinct studies provide specifications within each column. Specifications restricted to those with F-statistics less than 18. First two columns are further restricted to specifications where the usual t-ratio is valid because 0 is contained in the interval $[\hat{\rho}^{-1}(.565), \hat{\rho}^{-1}(-.565)]$.

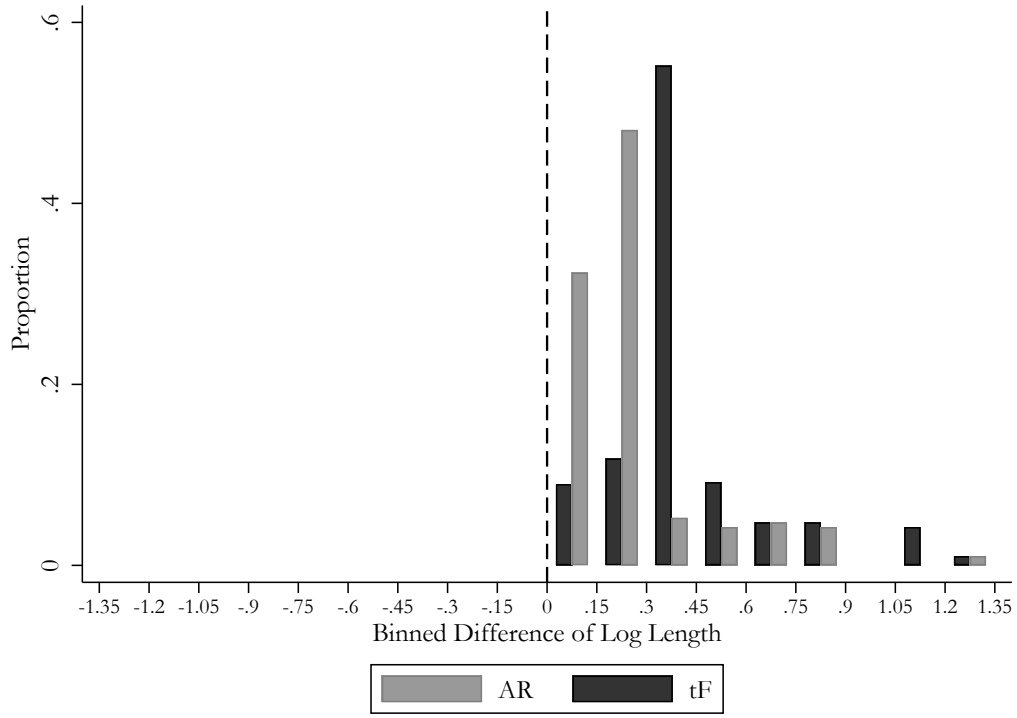
other two procedures that are valid without imposing any additional *a priori* bounds. For each specification, we can compute the difference in lengths of the various confidence intervals: $\ln(length_{AR}) - \ln(length_{VtF})$ and $\ln(length_{tF}) - \ln(length_{VtF})$.

This distribution across the 62 specifications is shown in Figure 2, which plots the (weighted) histogram of these difference measures. It is expected that VtF intervals will be shorter than that of tF due to the use of an additional statistic from the data. That gain in precision is quite substantial. For 79 percent of the specifications, tF confidence intervals are longer than those of VtF by more than 30 log points.

Finally, Figure 2 depicts the distribution of the AR-VtF differences in lengths across the 62 specifications. We find that VtF's confidence interval lengths are always shorter than AR interval lengths. For the $AR - VtF$ difference, the mode of this distribution is 15 to 30 log points with the difference being greater than 30 log points in about 20 percent of the specifications. As a basis of comparison, standard 95 percent confidence intervals are longer than 90 percent confidence intervals by about $\ln\left(\frac{1.96}{1.645}\right) \approx .18$ and 99 percent confidence intervals are longer than 95 percent confidence intervals by about $\ln\left(\frac{2.58}{1.96}\right) \approx .27$.

Although one must be cautious about making broad conclusions about the "representative IV analysis" from a small sample of published studies – which potentially over-samples studies with strong first-stage F-statistics or statistically significant results – the data nevertheless provides some evidence that in practice, when F-statistics are as small as they are in this sample, VtF

Figure 2: Frequency Distributions of $\ln(\frac{length_{tF}}{length_{VtF}})$, $\ln(\frac{length_{AR}}{length_{VtF}})$



Note: Uses the 62 specifications for which the first-stage F-statistic is smaller than 18. Frequencies are weighted by the inverse of the number of specifications in each study. Each bar denotes a frequency within an interval of difference in log-length of 0.15.

inferences will be materially different and more precise, with a higher likelihood of statistically rejecting null effects, compared to AR, tF, and even to the usual ± 1.96 procedure (when focusing on cases when the latter is valid). In section III.D, we present a systematic and extensive comparison of VtF to other procedures across a broad range of possible realizations of the data in the IV setting; the findings are quite consistent with the patterns we find in this small sample of empirical studies.

III Overview of VtF: Background, Definition, Confidence Interval Performance

In this section, we provide a high-level, non-technical summary of our main theoretical results. We provide context for the VtF procedure, and define and describe the test procedure and associated confidence set procedures in more detail. The section concludes with a systematic comparison of VtF, AR, and tF confidence interval performance. Details of the derivation of the VtF critical values and confidence intervals can be found in Appendix D.

III.A Context: the inferential problem with t-ratio inference and existing approaches

As has been understood for over two decades, the potential for weak instruments renders the conventional 2SLS t-ratio-based hypothesis test invalid, meaning the t-ratio can reject the null hypothesis far more than the intended significance level. That is, even when the null is true, and even in large samples, the test that rejects the null when $|\hat{t}| > 1.96$ can reject more than 5 percent of the time. The extent of over-rejection will depend on the true strength of the first-stage and the degree of endogeneity of the regressor X, both of which are unknown to the researcher, but the rejection rate can be arbitrarily close to 100 percent.

To see this more concretely, under the now-standard weak instrument asymptotics of Staiger and Stock (1997), the usual t-ratio $\hat{t}$ converges in distribution to a random variable which we will denote t, and the first-stage F-statistic, $\hat{F}$ analogously converges in distribution to a random variable which we will denote F. The random variable t (which can be characterized as a function of a bivariate normal vector) has a non-normal distribution that depends on two unknown quantities: the correlation of the main equation and first-stage errors, ρ, and the unit-free normalization of the first-stage coefficient, f_0.[25] These nuisance parameters ρ and f_0, in turn, determine the asymptotic rejection rate of the t-ratio under the null:

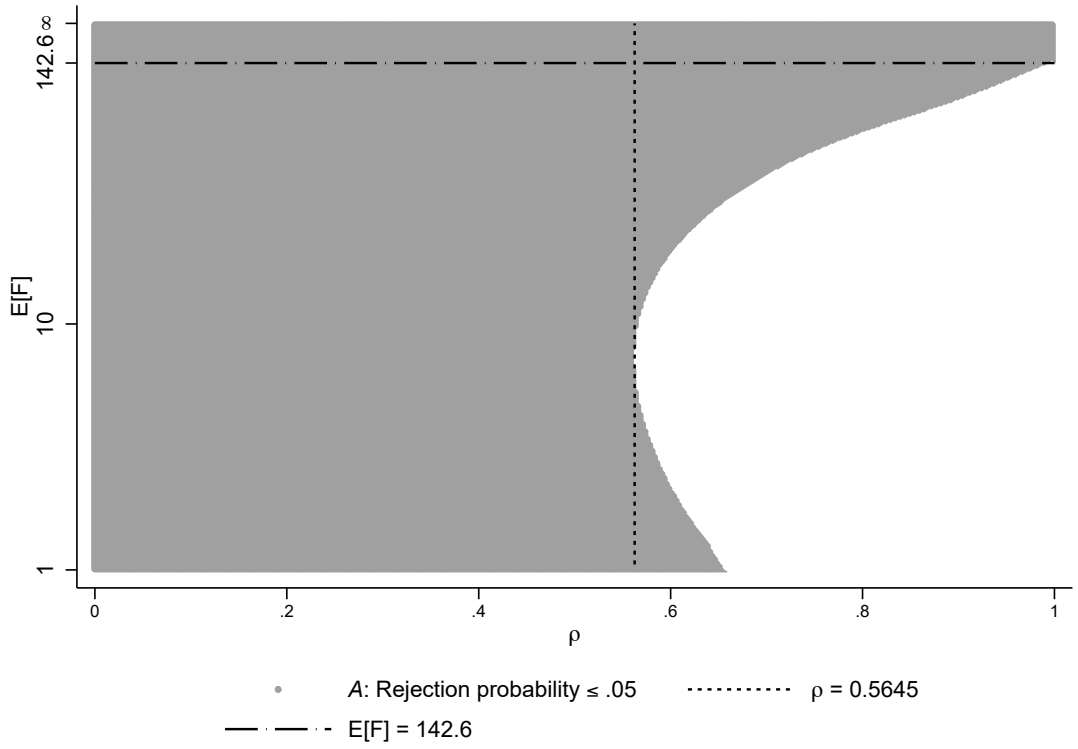
$$\Pr_{\rho,f_0}[|t| > 1.96] \leq 0.05 \quad \forall (\rho, f_0) \in A, \quad \text{and} \quad \Pr_{\rho,f_0}[|t| > 1.96] > 0.05 \quad \forall (\rho, f_0) \notin A, \tag{4}$$

where A is defined to be the set of ρ, f_0 values such that the usual t-ratio procedure will reject no more than the intended 5 percent of the time under the null hypothesis. The weak instrument over-rejection problem is encapsulated by the fact that A does not contain all ρ, f_0 values. The exact form of the set A directly follows from Staiger and Stock (1997) and SY (2005) and is recently visualized in both Angrist and Kolesár (2023) and Lee et al. (2020), and is replicated from the latter as the gray area in Figure 3 with $E[F] = {f_0}^2 + 1$. It is important to note that in addition to the anti-conservativeness that causes invalidity, the usual procedure can also be conservative. For ρ, f_0 values in the interior of the set A, the rejection rates are strictly *lower* than 0.05 (for a detailed contour plot, see Angrist and Kolesár (2023)).

The approach to just identified IV inference advocated in the econometric literature (see Andrews, Stock and Sun (2019), Keane and Neal (2023), Andrews, Moreira and Stock (2006)) is the *AR* test. This procedure avoids both the conservativeness and the anti-conservativeness of the usual

[25] Specifically, under the weak IV asymptotics of Staiger and Stock (1997), $f_0 \equiv \frac{\sqrt{N}\pi_N}{\sqrt{\sigma_{22}}}$ where π_N shrinks to zero at rate $\frac{1}{\sqrt{N}}$, and σ_{22} is the asymptotic variance of the first-stage coefficient estimator.

Figure 3: Values for ρ and $E[F]$ for which $|t| > 1.96$ is valid



t-test by achieving a rejection rate that matches the intended significance level for all values of the parameters ρ, f_0. *AR* uses information from the reduced-form residual covariance matrix to obtain estimates of σ_{11}, σ_{22}, and ρ_{RF} where

$$\sqrt{N}\begin{pmatrix} \widehat{\pi\beta} - \pi_N\beta \\ \hat{\pi} - \pi_N \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \rho_{RF}\sqrt{\sigma_{11}\sigma_{22}} \\ \rho_{RF}\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{22} \end{pmatrix}\right).$$

Here estimates of σ_{11}, σ_{22}, and ρ_{RF} could incorporate the use of a robust variance estimator (e.g. clustered, Newey-West, etc.) to match departures from homoskedasticity in the reduced-form errors. These estimates can then used to obtain the "t-ratio form" of the *AR* statistic:

$$\hat{t}_{AR} \equiv \frac{\hat{\pi}\left(\hat{\beta} - \beta_0\right)}{\frac{1}{\sqrt{N}}\sqrt{\hat{\sigma}_{11} - 2\beta_0\hat{\rho}_{RF}\sqrt{\hat{\sigma}_{11}\hat{\sigma}_{22}} + \beta_0^2\hat{\sigma}_{22}}}.$$

Under weak instrument asymptotics, $\hat{t}_{AR}$ converges in distribution to a random variable t_{AR}, which

has a standard normal distribution under the null. As a result, the *AR* test is asymptotically similar:

$$\Pr_{\rho,f_0}\left[|t_{AR}| > 1.96\right] = 0.05 \quad \forall(\rho, f_0)\,.$$

Given the *AR* procedure's ability to achieve correct size regardless of the value of $(\rho,\ f_0)$, it is considered to be "robust" to weak instruments and hence valid.

In practice, applied researchers typically do not employ *AR*.[26] Instead, as covered in numerous textbooks and surveys of the literature (e.g., Angrist and Pischke (2009), Stock and Watson (2019), Wooldridge (2019), or Hansen (2022)), they explicitly or implicitly recognize the problem of weak instruments by reporting the first-stage F-statistic with the typical motivation that the larger the F-statistic, the less concern they have with possible inferential distortions that arise from using the 1.96 rule. This standard practice in applied work is reflected in the over 9500 citations to SY (2005), which provides first-stage F-statistic thresholds to reduce inferential distortions.

When researchers take this approach, they are implicitly making inferences about β using the following rule: reject the null hypothesis if $\hat{t}^2 > 1.96^2$ *and* $\hat{F} > \bar{F}$, and otherwise accept the null hypothesis.[27] As before, weak instrument asymptotics will provide the limit distribution for $(\hat{t},\hat{F})$ which we will denote by (t,F). Let $\alpha_{\bar{F}} = \sup_{\rho,f_0}\Pr_{\rho,f_0}\left[\{|t| > 1.96\}\cap\{F > \bar{F}\}\right]$ denote the asymptotic size of this procedure for a given threshold $\bar{F}$. Typical implementations reduce the size distortion of the t-test but do not achieve the intended significance level associated, for example, with the value 1.96, i.e. $\alpha_{\bar{F}} > .05$. These tests translate to confidence intervals that take the form of $\hat{\beta} \pm 1.96 \cdot \hat{se}\left(\hat{\beta}\right)$ when $\hat{F} > \bar{F}$, and otherwise the *whole real line*, when $\hat{F} \leq \bar{F}$. These intervals have confidence levels equal to $1-\alpha_{\bar{F}}$, e.g. 88.6, 90.6, and 95 percent, for $\bar{F}$ equal to 10, 16.38, and 104.67, respectively. While convenient to implement, this approach can limit the applicability of instrumental variables, even when the first-stage might be considered to be quite strong but still below the threshold $\bar{F}$.

To achieve a 5/95 percent level of significance/confidence, while maintaining informative inference even when $\hat{F}$ is low, LMMP (2022) derive a refinement to this F-based approach: rather than a single-threshold rule, a smooth critical value function $c_{tF}(F)$ is used satisfying

$$\Pr_{\rho,f_0}\left[|t| > \sqrt{c_{tF}(F)}\right] \leq 0.05 \quad \forall(\rho, f_0)\,. \tag{5}$$

The function $c_{tF}(F)$ plateaus when $F > 104.67$ and asymptotes to infinity as F approaches 1.96^2 from above. LMMP (2022) shows that that there does not exist another critical value function

[26] LMMP (2022) report that it is used less than 4 percent of the time in their sample of *AER* articles.

[27] SY (2005) develop various thresholds in the context of testing the null hypothesis of the presence of weak instruments. Staiger and Stock (1997) discuss how to incorporate such a test for instrument weakness into inference on β.

uniformly below $c_{tF}(F)$ that also satisfies the condition for intended significance level (5).

While the tF procedure addresses the anti-conservativeness of the usual t-ratio approach, it does not address its conservativeness. As documented in LMMP (2022) and emphasized in Keane and Neal (2023), when the true ρ is small, it can have very low power for local alternatives, compared to AR, for example. This occurs because tF, like any test that achieves an intended significance level (e.g., the original Stock-Yogo approach), must accommodate all possible values of the parameters, including the worst-case/least-favorable scenario for rejection probabilities. In the case of this IV model, the worst-case is when ρ takes on extreme values such as ± 1.

More recently, Angrist and Kolesár (2023) introduce a novel strategy to justify the usual ± 1.96 t-ratio procedure. Instead of relying on the information contained in $\hat{F}$, they leverage the features of the set A defined by (4) and shown in Figure 3. They point out that there may be some situations in which it is reasonable to simply rule out, *a priori*, values of the parameters outside of the set A for which the usual t-test with critical value 1.96 is valid. Specifically, they use the fact that ρ is a one-to-one transformation of β via

$$\rho(\beta) = \frac{\rho_{RF} - \beta\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}}{\sqrt{1 - 2\rho_{RF}\beta\sqrt{\frac{\sigma_{22}}{\sigma_{11}}} + \left(\beta\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}\right)^2}}. \tag{6}$$

This transformation can effectively be treated as a known function, since $\hat{\sigma}_{11}$, $\hat{\sigma}_{22}$, and $\hat{\rho}_{RF}$ are all consistent estimators of σ_{11}, σ_{22}, and ρ_{RF}, respectively, even under weak-IV asymptotics. We write $\hat{\rho}(\cdot)$ when consistent estimators of the covariance parameters are used in (6). [28]

Angrist and Kolesár (2023) emphasize the fact that any hypothesis about β is thus equivalent to a specific hypothesis about ρ and vice versa, and inference about β maps directly to inference about ρ. Thus, if one is comfortable providing *a priori* bounds for a range of reasonable values of β, this implies corresponding bounds on ρ. If those bounds imply $|\rho|$ is less than .565, then (ρ, f_0) will be in the set A and the usual ± 1.96 intervals will remain valid at the 95 percent confidence level. Analogously, making assumptions about β such that $|\rho|$ is always less than .435 would allow the usual ± 2.58 intervals to be valid at the 99 percent confidence level. Angrist and Kolesár (2023) find that in three prominent empirical examples, their bounds for β would imply values of $|\rho|$ less than .565, establishing that conventional t-ratio inference is justified in those three cases.

The main limitation to this approach is that it requires the researcher to 1) rule out some values of β on *a priori* grounds, leading to $\beta \in [\beta_{lower}, \beta_{upper}]$, 2) ensure that the remaining allowable values of $\beta \in [\beta_{lower}, \beta_{upper}]$ are contained within a region such that $|\hat{\rho}(\beta)| \leq .565$, and 3) be willing to report confidence sets of the form $[\hat{\beta} \pm 1.96 \cdot \hat{se}(\hat{\beta})] \cap [\beta_{lower}, \beta_{upper}]$ (i.e., the usual

[28] Note that under any null hypothesis that $\beta = \beta_0$, $\hat{\rho}(\beta_0)$ is a consistent estimator of ρ.

confidence interval that is truncated by the *a priori* bounds β_{lower} or β_{upper}; or possibly the empty set).

This paper proposes a procedure aimed at achieving a common goal of both the strategies of SY (2005) and Angrist and Kolesár (2023): to allow the researcher to use the usual ± 1.96 confidence intervals – whenever possible. It does so, by further enhancing the standard F-based critical value function approach of Stock-Yogo, addressing the anti-conservativeness of the usual ± 1.96 procedure, while also addressing the conservativeness of tF. In doing so, we provide a confidence interval procedure that does not require imposing assumptions about the parameter β, and more generally a test procedure for *any* hypothesized value of β, applicable to *any* dataset.

We are motivated by a desire to offer an improvement that resembles already-prevailing practice ("t-ratio plus F-statistic" inference), and so a natural question is what price in terms of power one must pay for that convenience and resemblance with existing practice. As we show below in Section III.D, our comparisons of VtF confidence intervals to that of AR reveal that there is apparently virtually no price to pay. To the contrary, VtF confidence intervals generally outperform those of tF, AR, and in some realizations of the data, even the usual (and invalid) ± 1.96 intervals.

III.B Description of VtF test procedure

As discussed in the previous section, we seek a test procedure that: 1) uses the conventional and familiar 2SLS t-ratio 2) uses a critical value function that, given a fixed null value of β_0 (and hence fixed null value of $\hat{\rho}(\beta_0)$ via the sample version of Equation (6)), depends only on the first stage $\hat{F}$-statistic, and thus 3) avoids conservativeness, by adapting to the correlation parameter implied by the null.

Assuming, for the moment, the existence of such a critical value function, these properties could be formally expressed as a test that rejects if and only if

$$|\hat{t}| > \sqrt{c\left(\hat{\rho}(\beta_0), \hat{F}; \alpha\right)}$$

where the critical value function $c(\cdot,\cdot;\alpha)$ satisfies the condition:

$$\text{Pr}_{\rho, f_0}\left[|t| > \sqrt{c\left(\rho(\beta_0), F; \alpha\right)}\right] = \alpha, \quad \forall \rho, f_0 \tag{7}$$

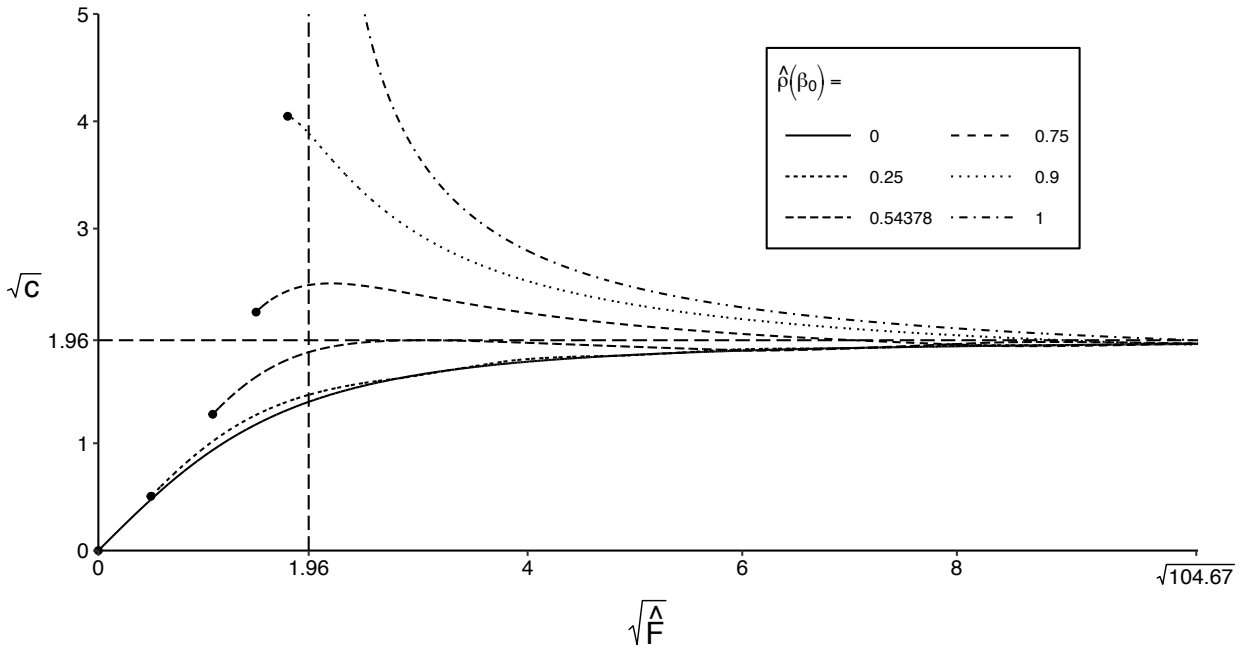
under the null.[29] In our discussion, we focus on the case of $\alpha = 0.05$, but, to be clear, VtF can accommodate any size α (e.g., $\alpha = 0.01$).

[29]Note that under the null, $\rho(\beta_0) = \rho(\beta) = \rho$. A statement equivalent to (7) that explicitly acknowledges the one-to-one relation between β and ρ given the reduced-form covariance matrix is $\text{Pr}_{\beta_0, f_0, \sigma_{11}, \sigma_{22}, \rho_{RF}}\left[|t| > \sqrt{c\left(\rho(\beta_0), F; 0.05\right)}\right] = 0.05 \ \forall \beta_0, f_0, \sigma_{11}, \sigma_{22}, \rho_{RF}$.

Note that the equal sign in (7) requires that the test not only controls size, but that it is also a similar test (so that it avoids being overly conservative for any parameter values).

While the statement in Equation (7) is simple, establishing whether such a function $c(\cdot,\cdot;\alpha)$ exists is not. In Appendix D, we establish that such a function $c(\cdot,\cdot;\alpha)$ exists and also establish that it is unique (within a class of candidate functions).

Figure 4: VtF Critical Value function



We now illustrate in Figure 4 the VtF critical value function $\sqrt{c\left(\hat{\rho}(\beta_0),\hat{F};\alpha\right)}$ for selected values of $\hat{\rho}(\beta_0)$ and for $\alpha = 0.05$, noting the following features:

1. For any value of $\hat{\rho}(\beta_0)$, as F increases, the critical value function $\sqrt{c\left(\hat{\rho}(\beta_0),\hat{F}\right)}$ tends to the usual standard normal critical value of 1.96.
2. When the null value implies a high degree of endogeneity, i.e., $|\hat{\rho}(\beta_0)|$ is far enough from zero, the critical value function is decreasing in $\hat{F}$. As $|\hat{\rho}(\beta_0)|$ approaches 1, it approaches the tF critical value function of LMMP (2022), which is strictly decreasing in $\hat{F}$.
3. By contrast, when the null value implies a low degree of endogeneity, the critical value function is increasing in $\hat{F}$. As $|\hat{\rho}(\beta_0)|$ approaches zero, $c\left(\hat{\rho}(\beta_0),\hat{F}\right)$ converges to $\frac{1.96^2}{1+1.96^2/\hat{F}}$, which is strictly increasing in $\hat{F}$. Moreover, when the null corresponds to $\hat{\rho}(\beta_0) = 0$, the VtF test is identical to the AR test. See Appendix D for details on both of these points.
4. When $|\hat{\rho}(\beta_0)| \leq 0.543$, the critical value function is entirely below the value 1.96.

5. For any null with $|\hat{\rho}(\beta_0)| < 1$, VtF may reject the null even when the first stage coefficient is statistically insignificant (i.e., $\hat{F} < 1.96^2$). However, it can be shown that whenever $\hat{F} < 1.96^2$, it is impossible to reject the null that $\beta_0 = \pm\infty$, i.e., VtF confidence sets are unbounded. This condition for an unbounded confidence set is identical to that of AR.

6. For each value of $\hat{\rho}(\beta_0)$, the critical value function $c(\hat{\rho}(\beta_0), \cdot)$ is drawn in Figure 4 to start at the point $\left(\hat{\rho}^2(\beta_0)1.96^2, \frac{\hat{\rho}^2(\beta_0)1.96^2}{1-\hat{\rho}^2(\beta_0)}\right)$. At this point, the critical value function intersects the function $\frac{\hat{F}}{1-\hat{\rho}^2(\beta_0)}$ which is an upper bound of $\hat{t}^2$ for $\hat{\rho}(\beta_0)$. For $\hat{F} \in [0, \hat{\rho}^2(\beta_0))$, the VtF test will always accept, which can be achieved by setting the critical value function c to any value above the upper bound, e.g. for $\hat{F} \in [0, \hat{\rho}^2(\beta_0))$, $c\left(\hat{\rho}(\beta_0), \hat{F}\right) = \infty$ or $\frac{\hat{\rho}^2(\beta_0)(1.96)^2}{1-\hat{\rho}^2(\beta_0)}$. Since any function above the upper bound $\frac{\hat{F}}{1-\hat{\rho}^2(\beta_0)}$ will equivalently lead to the same test, we do not draw in this part of the critical value function.

Appendix Table A3 provides tables of VtF critical values for selected values of $\hat{\rho}(\beta_0)$ and for $\alpha = 0.05$ and $\alpha = 0.01$. For intermediate values of $\hat{F}$ and $\hat{\rho}(\beta_0)$, linear interpolation using these tables can provide a rough guide; critical values for any values of $\hat{F}$ and $\hat{\rho}(\beta_0)$ can be obtained via STATA code provided at http://www.princeton.edu/~davidlee/wp/SupplementVtF.html. Since these critical values are a function of only two variables ($\hat{F}$ and $\hat{\rho}(\beta_0)$), with an appropriate flexible functional form, one could potentially compactly represent a good approximation to the critical value function using those two variables and a relatively small number of parameters.

Remark. There are two immediate and important implications of the fourth point above, which notes when the VtF critical value function lies below 1.96. Since the critical value function is constructed to produce a probability of rejection of 0.05 under the null, the first consequence of Point 4 is that for any for any value β_0 such that $|\rho(\beta_0)| \leq 0.543$, the null rejection probability for the usual ± 1.96 test must be lower than 0.05. Thus, since power (rejection probability under alternative values) is continuous in the parameters of the model for both procedures, it must be true that VtF will be more powerful than the ± 1.96 procedure for any "local" alternative value of β. Also note that the condition $|\rho(\beta_0)| \leq .543$ is close to the "valid zone" $|\rho(\beta_0)| \leq .565$ shown in Figure 3. This means that when we restrict attention to hypotheses for which the usual ± 1.96 procedure is valid (as we describe in Section II; see Angrist and Kolesár (2023)), we would expect VtF to be more powerful in the above sense. This expectation is consistent with our findings that VtF was more successful than the usual ± 1.96 procedure at statistically detecting null effects in our sample of studies as reported in Section II. Notably, by contrast, both single-threshold F-based procedures and tF, for these restricted null values, have inferior power compared to the ± 1.96 procedure.

Second, the same principle applies to the confidence interval, which is by definition the set of hypothesized values that are accepted by the corresponding test procedure, for a given realization of

the data. Under the same conditions necessary to justify the validity of the usual ± 1.96 intervals, VtF confidence intervals will tend to be shorter. Suppose, for example, one were comfortable imposing a condition that would ensure validity of ± 1.96 intervals – that all β within the interval $[\beta_{lower}, \beta_{upper}]$ satisfied $|\hat{\rho}(\beta)| \leq 0.543$. Due to Point 4 above, any such hypothesized value that is rejected by the rule $|\hat{t}| > 1.96$ will always be rejected by the VtF critical value, but the converse statement is not true. Thus, in this case, VtF confidence intervals will be entirely contained within the ± 1.96 intervals for this range of β.

In sum, although VtF does not require any restriction on the range of hypotheses allowable to test, imposing those restrictions will, in any case, result in more precise inferences than what would be produced by the usual ± 1.96 procedure.

III.C VtF Confidence Intervals

A confidence set can be defined as the set of hypothesized values that would be accepted by a given test procedure. In Appendix D, noting that the function $\hat{\rho}(\cdot)$ is determined by $\hat{\sigma}_{11}, \hat{\sigma}_{11}$, and $\hat{\rho}_{RF}$, we are able to show VtF confidence set can be written as the set-valued function that depends on the quantities $\hat{\beta}, \hat{se}(\hat{\beta}), \hat{r}, \hat{F}$:

$$CS\left(\hat{\beta}, \hat{se}(\hat{\beta}), \hat{r}, \hat{F}\right) = \left\{ \beta_0 \,\middle|\, -\sqrt{c\left(\hat{\rho}(\beta_0), \hat{F}\right)} \leq \frac{\hat{\beta} - \beta_0}{\hat{se}(\hat{\beta})} \leq \sqrt{c\left(\hat{\rho}(\beta_0), \hat{F}\right)} \right\}$$

where $\hat{r} = \hat{\rho}(\hat{\beta})$. This confidence set has the property that

$$\liminf \Pr_{\rho(\beta), f_0}\left[\beta \in CS\left(\hat{\beta}, \hat{se}\left(\hat{\beta}\right), \hat{r}, \hat{F}\right)\right] = .95$$

and therefore has correct confidence level and is also never conservative meaning the liminf in the above equation could be replaced by limsup and the same equality would still hold.

As we explain in greater detail in Appendix D, when it is bounded, the confidence set can always be contained by a single interval conveniently represented as

$$\left[\hat{\beta} - k^{-}\left(\hat{r}, \hat{F}\right) \cdot \hat{se}\left(\hat{\beta}\right), \hat{\beta} + k^{+}\left(\hat{r}, \hat{F}\right) \cdot \hat{se}\left(\hat{\beta}\right)\right] \tag{8}$$

where $k^{-}(\cdot,\cdot)$ and $k^{+}(\cdot,\cdot)$ are functions used to construct the endpoints of the confidence interval. Recall that the usual (and invalid) approach to confidence intervals for the IV parameter is to inflate the standard error by 1.96. The VtF approach to confidence interval construction is to replace 1.96 with the data-dependent factors $k^{-}(\cdot,\cdot)$ and $k^{+}(\cdot,\cdot)$ that lead to correct coverage rates. These data-dependent factors can be greater or less than 1.96. To economize on notation, and hopefully at no

risk of confusion, we write k^- in place of $k^-(\cdot,\cdot)$ and k^+ in place of $k^+(\cdot,\cdot)$.

It is important to note that because the factors k^- and k^+ are not the same, VtF confidence intervals will generally be asymmetric, so that the length of the upper segment (the upper bound minus $\hat{\beta}$) will be different from that of the lower segment ($\hat{\beta}$ minus the lower bound). Which segment is longer turns out to depend on the sign of $\hat{r}$. In particular, if $\hat{r}$ is positive, then the upper segment of the confidence interval will be shorter than the lower segment; and if $\hat{r}$ is negative, then the upper segment of the confidence interval will be longer.

Remark. Practitioners are most familiar with reporting two numbers, the 2SLS point estimate and standard error, with the latter displayed right below the former. From these numbers, it is easy for the reader to instantly approximate the usual ± 1.96 confidence intervals. If this reporting convenience is important, it would be straightforward to report a (conservative) "adjusted" standard error, by multiplying $\frac{\max[k^+(\hat{r},\hat{F};0.05),k^-(\hat{r},\hat{F};0.05)]}{1.96}$ by $\hat{se}(\hat{\beta})$. This could be reported as a "symmetric VtF 0.05 standard error".[30] Note that this reporting practice would lead to conservative inferences and the resulting symmetric-around-$\hat{\beta}$ interval would be unnecessarily longer than the unsymmetrized interval, achieving coverage that exceeds e.g. 95 percent. An alternative – just as in the case for AR – is to simply report the VtF interval (8) in addition to $\hat{\beta}$ and $\hat{se}(\hat{\beta})$. Presuming that the $\hat{F}$-statistic is already reported, the additional reporting of $\hat{r}$ would then provide full transparency, and give any reader the ability to re-construct intervals for any of the methods mentioned in this paper.

Appendix Table A5 presents k^- and k^+ for selected values of $|\hat{r}|$ and $\hat{F}$. The tables reveal the following broad patterns: 1) There is a wide range of possible inflation factors, ranging from values of k^+ and k^- that are *smaller* than the usual inflation factor of 1.96, to values larger than 5; 2) irrespective of the value of $|\hat{r}|$, k^+ and k^- tend to 1.96 as $\hat{F}$ increases; 3) k^+ and k^- are generally decreasing in $\hat{F}$, but there are some regions where they are slightly increasing in $\hat{F}$, for example when $k^+, k^- < 1.96$; and 4) for any fixed $\hat{F}$ and $\hat{r} > 0$, k^- is monotonically increasing in $\hat{r} > 0$ while k^+ is non-monotonic in $\hat{r} > 0$ (and the symmetric relationships apply when $\hat{r} < 0$).

Finally, the entries where both k^+ and k^- are below 1.96 are shaded in gray in Appendix Table A5. For these values of $\hat{F}$ and $|\hat{r}|$, the practitioner could elect to forgo the smaller VtF interval and simply use the conventional t-ratio confidence intervals and still obtain valid (even if conservative) inference. The pattern of shaded entries shows that even when first-stage F-statistics are relatively low (e.g., $\hat{F}$ of 20), one could adopt the usual ± 1.96 intervals provided $|\hat{r}|$ is sufficiently low. Otherwise, the practitioner can simply use the inflation factors k^+ and k^-. A simple (conservative) rule of thumb for this region, in which one can revert to the usual ± 1.96 intervals is given by the condition $\hat{F} > 10 + 100\,|\hat{r}|$.

[30] Analogously, $\frac{\max[k^+(\hat{r},\hat{F};0.01),k^-(\hat{r},\hat{F};0.01)}{2.58} \cdot \hat{se}(\hat{\beta})$ with the analogous k^+, k^- factors for the 99 percent levels, could be reported as the "symmetric VtF 0.01 standard error".

We again note that for intermediate values of $\hat{F}$ and $\hat{r}$, linear interpolation using these appendix tables can provide a rough guide; inflation factors for any values of $\hat{F}$ and $\hat{r}$ can be obtained via STATA code provided at http://www.princeton.edu/~davidlee/wp/SupplementVtF.html. Since these inflation factors are a function of only two variables ($\hat{F}$ and $\hat{r}$), with an appropriate flexible functional form, one could potentially compactly represent a good approximation to the inflation factors using those two variables and a relatively small number of parameters.

III.D Confidence Interval Length Performance: VtF versus t, tF and AR

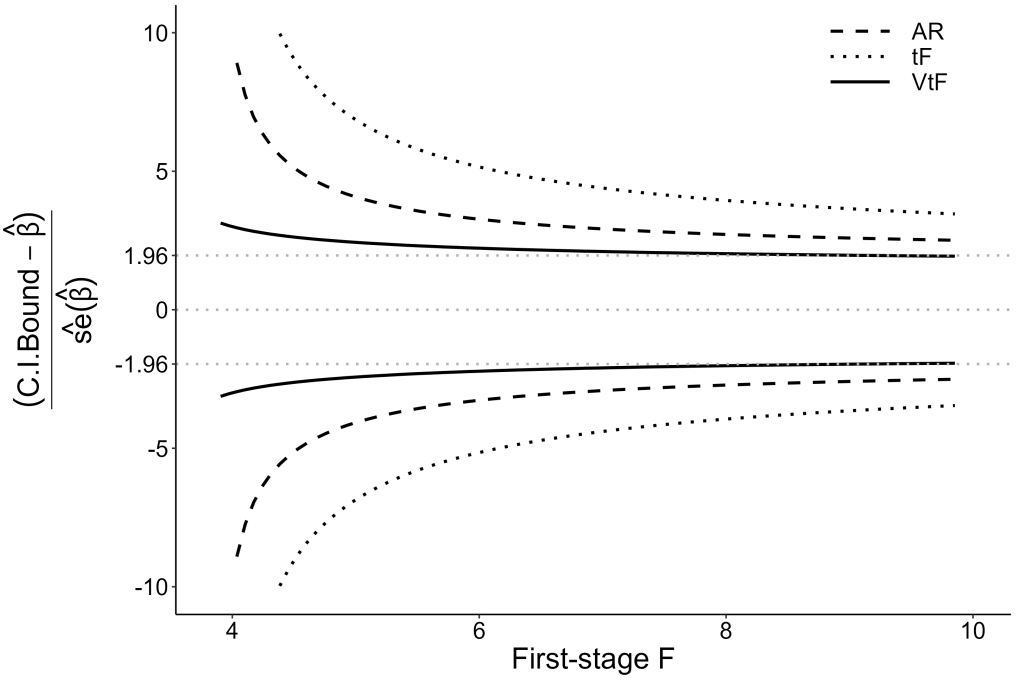
In Section II, VtF confidence intervals were compared to other confidence interval methods for a collection of empirical studies. In this subsection, we supplement that presentation with a comprehensive and systematic comparison that is not tied to any particular sample of empirical studies. We find that VtF produces significantly shorter confidence intervals than that of tF, which demonstrates that there is considerable value in using the information contained in the statistic $\hat{r}$. We also find that VtF intervals are generally shorter than those of AR – roughly shorter to about the same extent that AR is shorter than tF; this superior performance necessarily must stem from the structure of the test (VtF is a Wald-based test, while AR is a Lagrange Multiplier or Likelihood Ratio test), since both the VtF and AR tests use the same information: $\widehat{\pi\beta}, \hat{\pi}, \hat{\sigma}_{11}, \hat{\sigma}_{22}, \hat{\rho}_{RF}$. We also systematically document the possibility that VtF intervals can be shorter than the conventional (and invalid) ± 1.96 intervals, even when the VtF intervals are not entirely contained within the ± 1.96 intervals. By contrast, AR intervals are never shorter than the usual ± 1.96 confidence intervals (and therefore could never be entirely contained within the usual confidence intervals).

Before proceeding, it is important to note exactly what is entailed in the comparisons made here and described above. Typically statistical procedures are compared by looking at features of their outcome distributions, e.g., bias, variance, and power. This kind of comparison is typically made for specific data generating processes. We provide this more standard power analysis, as well as an analysis of the *distribution* of confidence interval lengths for 16 different data generating processes, corresponding to $\rho = 0, 0.5, 0.8, 0.9$ and $f_0 = 1, 3, 6, 9$ in Section IV, and Appendices C.1 and C.2.

By contrast, the comparison in this section does not involve distributions, expectations, or data generating processes and is instead done at the much finer level of data realizations. Above we noted that the VtF inflation factors k^- and k^+ depend on the data only through the values of $\hat{r}$ and $\hat{F}$. We denote the analogous inflation factors for AR and tF by k^-_{AR}, k^+_{AR}, k^-_{tF} and k^+_{tF}. Since these inflation factors also depend on the data only through $\hat{r}$ and $\hat{F}$, we can make a complete set of relative length comparisons for all possible data realizations by considering just the comparisons for each $\hat{r}$ and $\hat{F}$. Any statement about the distribution of lengths in repeated samples for a given data generating process would be driven by the implied distribution of $(\hat{r}, \hat{F})$.

Figure 5: Confidence Interval Factors, VtF, tF, AR, t

(a) $\hat{r} = 0$



(b) $\hat{r} = -0.5$

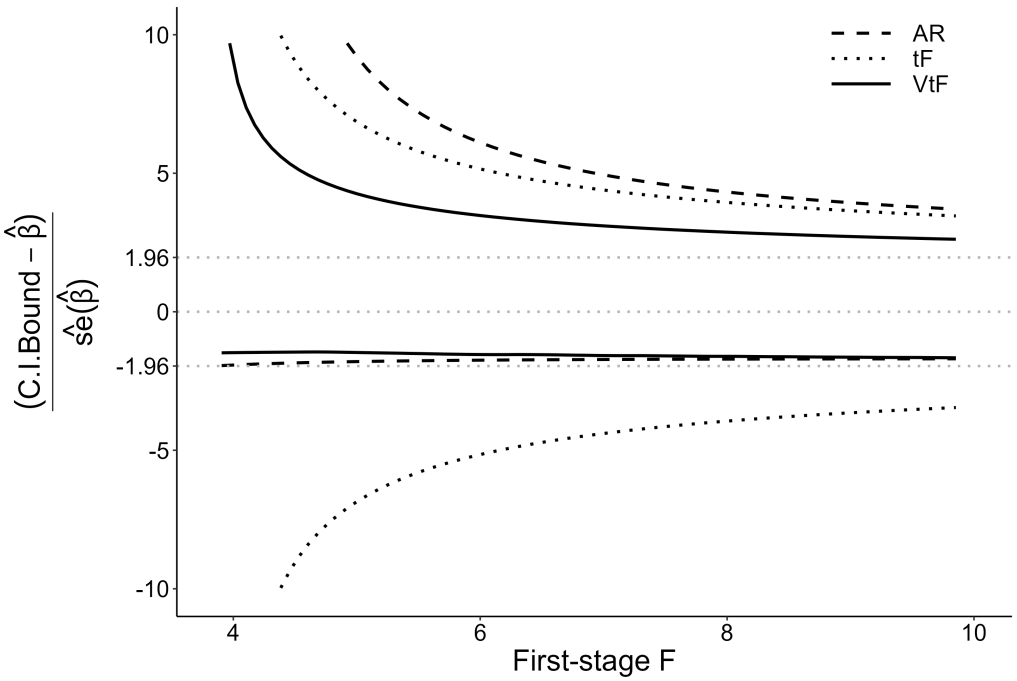


Figure 5a plots the inflation factors k^- to k^+ at a finer level of resolution than the corresponding table in Appendix Table A5 for the case that $\hat{r} = 0$. In addition, we overlay the analogous inflation factors for tF and AR, where we focus on the region where $\hat{F} < 10$. The VtF intervals are clearly much shorter than those of tF, unsurprising since, compared to VtF or AR, the tF procedure essentially ignores the information in $\hat{r}$.

The VtF confidence intervals are also clearly shorter than (and contained within) that of AR, whose inflation factors are roughly midway between those of VtF and tF.

Figure 5b plots the inflation factors for the case that $\hat{r} = -0.5$. The figure shows how both VtF and AR produce asymmetric intervals around $\hat{\beta}$, and are longer on the upper segment when $\hat{r}$ is negative. Again, VtF and AR intervals substantially improve on the (symmetric) tF intervals, particularly on the lower segment of the interval. The lower bound of the VtF and AR intervals are quite similar and also above -1.96. As for the upper segment of the interval, we see again that VtF outperforms AR to a similar extent that AR outperforms tF.

Figures 6a-c provide another visualization of relative lengths for the all of the procedures for the range $\hat{F} \in (1.96^2, 104.67)$, and $0 \leq |\hat{r}| \leq 1$. In all of the figures, we additionally overlay the realizations of $\hat{F}$ and $\hat{r}$ from the 89 specifications such that $\hat{F} \in (3.84, 104.67)$ and use a (units-free) difference in the logs of length measure, for example,

$$\ln\left(\frac{\text{length}_{tF}}{\text{length}_{VtF}}\right) = \ln\left(k^+_{tF} + k^-_{tF}\right) - \ln\left(k^+ + k^-\right).$$

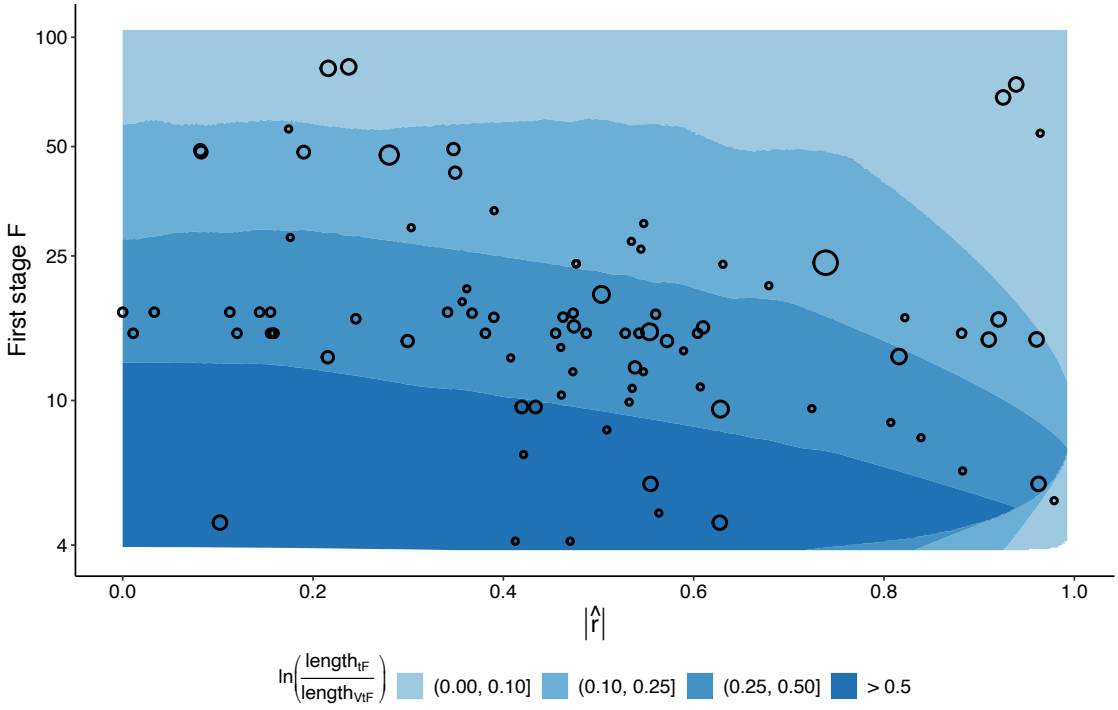
Figure 6a depicts four different ranges for the difference in log of lengths, $\ln\left(\frac{\text{length}_{tF}}{\text{length}_{VtF}}\right)$: $(0, .1]$, $(.1, .25]$, $(.25, .5]$, and $> .5$; and the values of $(|\hat{r}|, \hat{F})$ that lead to these lengths. Since VtF intervals are always shorter than tF intervals the difference in log lengths is always positive. The figure illustrates that in general as the first-stage F becomes smaller, VtF intervals become shorter than those of tF, and this is generally true for values of $|\hat{r}|$ less than 0.7. Above 0.7, the overall advantage of VtF relative to tF is still present but diminishes.

Figure 6b is analogous to Figure 6a, but represents the ranges of values for $\ln\left(\frac{\text{length}_{AR}}{\text{length}_{VtF}}\right)$, adding a fifth category, when $\ln\left(\frac{\text{length}_{AR}}{\text{length}_{VtF}}\right) < 0$. Overall, for most of this space, AR intervals are longer than VtF intervals, and the pattern of how relative length varies with $\hat{F}$ and $|\hat{r}|$ is similar to that depicted in Figure 6a, but it is clear that the VtF advantage over AR is not as great as the VtF advantage over tF. There is a relatively small region on the right side of the graph that represents realizations for which AR intervals are shorter. In probabilistic terms, this region on the far right of the graph is generally small. In Appendix C.1, under the 16 designs considered, the probability that the AR interval is shorter than the VtF interval is never greater than three percent.

Figure 6c provides another depiction of the comparison between VtF and AR intervals, using

Figure 6: Confidence Interval Lengths

(a) VtF versus tF



Note: Circles represent 89 specifications for which first-stage F-statistics are between 1.96^2 and 104.67, with the size of circle proportional to the inverse of the number of specifications in each study. The degree of shading indicates the difference in the log(CI lengths) (tF minus VtF). The vertical axis is a log-scale. The range of $|\hat{r}|$ is $[0, 0.995]$.

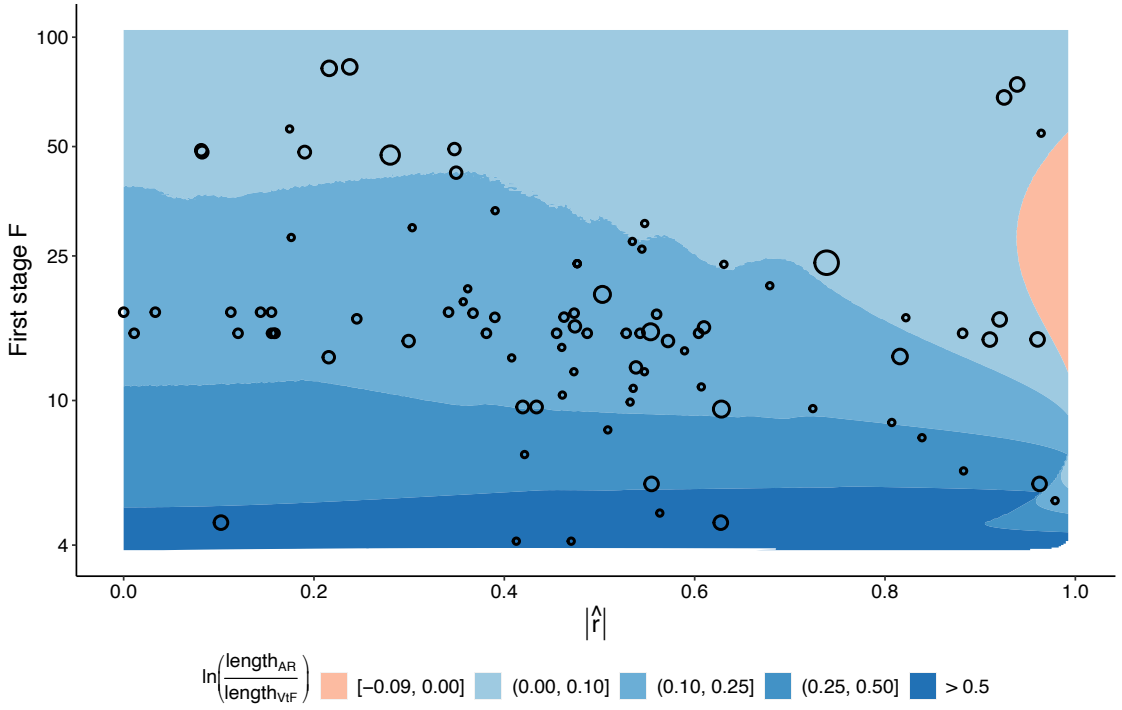
the following mutually exclusive categories: 1) $[-k^-, k^+] \subseteq [-k^-_{AR}, k^+_{AR}]$ (and hence $\ln\left(\frac{\text{length}_{AR}}{\text{length}_{VtF}}\right) > 0$); 2) $\ln\left(\frac{\text{length}_{AR}}{\text{length}_{VtF}}\right) > 0$ and $[-k^-, k^+] \not\subseteq [-k^-_{AR}, k^+_{AR}]$; and 3) $\ln\left(\frac{\text{length}_{AR}}{\text{length}_{VtF}}\right) < 0$ and $[-k^-, k^+] \not\supseteq [-k^-_{AR}, k^+_{AR}]$; there were no realizations for which $[-k^-, k^+] \supseteq [-k^-_{AR}, k^+_{AR}]$. For a substantial region (the left side of the graph), VtF intervals are entirely contained within AR intervals. Most of the remaining region represents data realizations for which VtF intervals are not entirely contained within AR intervals, but are nevertheless shorter.

Finally, Figure 6d is an analogous figure, comparing VtF and the conventional ± 1.96 intervals. Even though the ± 1.96 interval procedure does not have correct confidence level, we make the comparison to illustrate when it will be the case that VtF intervals are shorter than the usual ± 1.96 intervals, and in particular, when it would be clear that simply reverting to the ± 1.96 intervals would not affect the validity of the inference.

The figure shows that there is a substantial region where VtF intervals are longer (shades of red) than the ± 1.96 intervals. We know such a region must exist due to the under-coverage of

Figure 6: Confidence Interval Lengths

(b) VtF versus AR



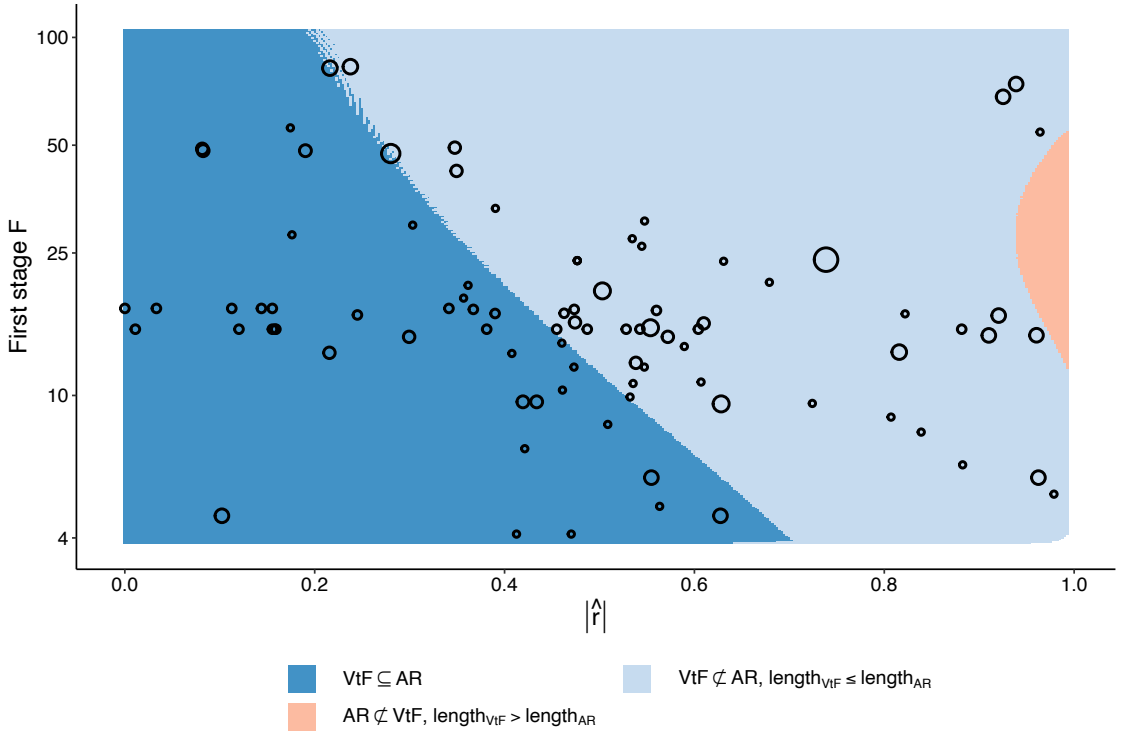
Note: Circles represent 89 specifications for which first-stage F-statistics are between 1.96^2 and 104.67, with the size of circle proportional to the inverse of the number of specifications in each study. The degree of shading indicates the difference in the log(CI lengths) (AR minus VtF). The vertical axis is a log-scale. The range of $|\hat{r}|$ is $[0, 0.995]$.

the ± 1.96 intervals. At the same time, the figure shows that there is a substantial region of larger $\hat{F}$ and/or small $|\hat{r}|$ for which the VtF intervals will be shorter (shades of blue). Among the 89 specifications from our sample of studies, 32 percent fall within this region. Within that region, there exists a substantial region for which the VtF intervals are entirely contained within the ± 1.96 interval (category 1). It is this last region that is conservatively approximated by the rule of thumb $\hat{F} > 10 + 100\,|\hat{r}|$, which is also depicted in the figure. The researcher can simply use the usual ± 1.96 intervals whenever $\hat{F} > 10 + 100\,|\hat{r}|$, and otherwise use the VtF intervals via Appendix Table A5, and be assured that the confidence level of 95 percent is uncompromised. It is important to note that a similar strategy cannot be used with AR intervals, which are always longer than the ± 1.96 intervals.

Not reported here (but available upon request) we produced a parallel set of heatmaps that compare the performance of a (conservative) symmetric version of the VtF interval formed by $\hat{\beta} \pm \max[k^+, k^-] \cdot \hat{se}(\hat{\beta})$, to tF, an analogously symmetrized AR, and the usual ± 1.96 intervals. The patterns are qualitatively similar, with the main difference being that the symmetrized AR

Figure 6: Confidence Interval Lengths

(c) VtF versus AR



Note: Circles represent 89 specifications for which first-stage F-statistics are between 1.96^2 and 104.67, with the size of circle proportional to the inverse of the number of specifications in each study. The shading indicates whether the VtF intervals contains/are contained by AR intervals, and if not, which interval is longer. The vertical axis is a log-scale. The range of $|\hat{r}|$ is $[0, 0.995]$.

intervals are longer than the symmetrized VtF intervals for all realizations of the data.

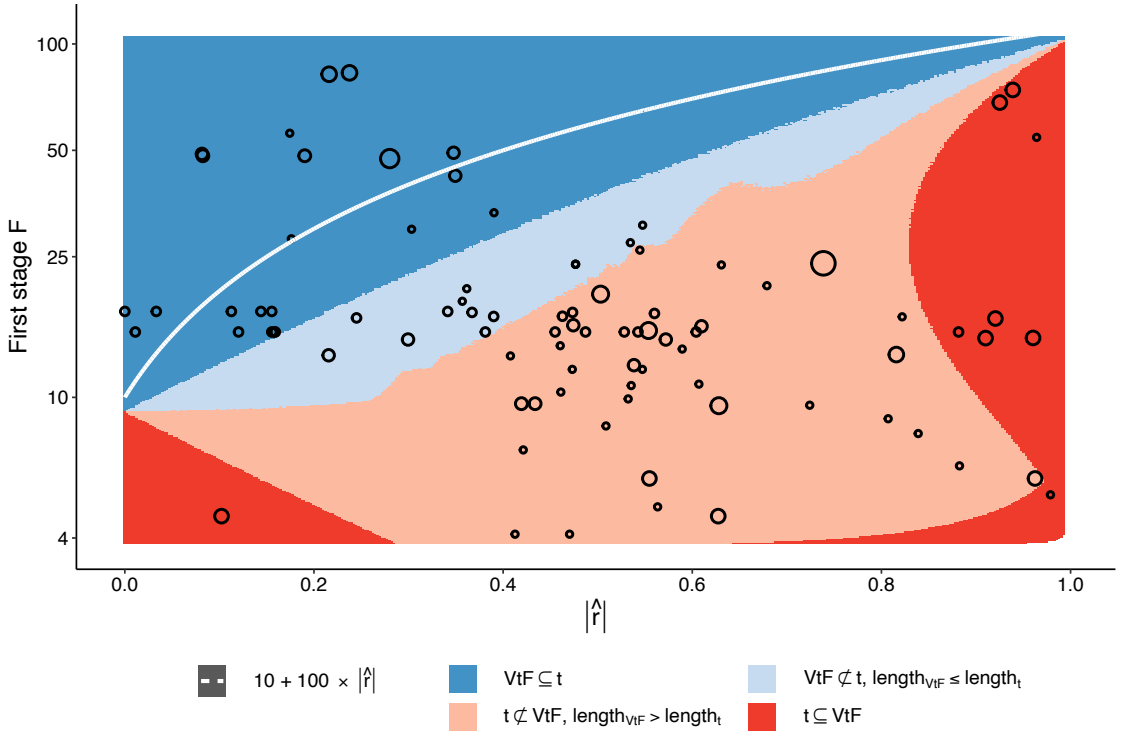
IV Power Comparisons with existing Methods

In previous sections, we presented a comprehensive comparison of VtF, AR, and tF procedures in terms of confidence interval lengths through a data realization by data realization evaluation. In our sample of empirical studies, VtF confidence intervals are shorter than both AR and tF intervals in 100 percent of the specifications. The heatmaps in Figure 6 provide a more systematic summary of confidence interval lengths for each possible data realization and confirm the broad scope of VtF's advantage.

We now conduct a more standard power analysis of these different testing procedures. Given the confidence interval length findings, we should expect to find that VtF has some power advantages relative to AR and tF. Any observed advantage may be somewhat surprising or appear to be at odds with two types of well established optimality results for AR in the case of the just-identified

Figure 6: Confidence Interval Lengths

(d) VtF versus t



Note: Circles represent 89 specifications for which first-stage F-statistics are between 1.96^2 and 104.67, with the size of circle proportional to the inverse of the number of specifications in each study. The shading indicates whether the VtF intervals contains/are contained by the usual ± 1.96 intervals, and if not, which interval is longer. The vertical axis is a log-scale. The range of $|\hat{r}|$ is $[0, 0.995]$.

IV model: a) AR is uniformly most powerful within a class of tests; and b) AR is an admissible procedure.[31] Below we present the power of these tests with the further aim of reconciling our findings with the optimality results from the previous literature. Additionally, we provide a new decomposition of power that clarifies how VtF focuses power differently, compared to AR, to achieve shorter confidence intervals.

IV.A The Weak IV Probability Model

To derive the VtF procedure (see Appendix D) and to conduct the power analysis, we use the now standard IV asymptotic framework of Staiger and Stock (1997), which provides accurate approximations regardless of the quality of the instrument. Using the same notation as in the

[31] See Andrews, Stock and Sun (2019) for a discussion of the literature's optimality findings on AR with respect to the just-identified model.

Online Appendix to LMMP (2022), $\hat{t}_{AR}(\beta_0)$ and $\hat{f}$ converge jointly in distribution to

$$(9) \qquad \begin{pmatrix} t_{AR}(\beta_0) \\ f \end{pmatrix} \sim N\left(\begin{pmatrix} f_0 \frac{\Delta(\beta_0)}{\sqrt{1+2\rho\Delta(\beta_0)+\Delta^2(\beta_0)}} \\ f_0 \end{pmatrix}, \begin{pmatrix} 1 & \rho(\beta_0) \\ \rho(\beta_0) & 1 \end{pmatrix} \right)$$

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

and ρ is the true population correlation between Zu and Zv. Equations (6) and (9) provide equivalent expressions for $\rho(\beta_0)$.[32] Equation (6) can be used to map any hypothesized value β_0 to the implied correlation between Zv and Zu under that hypothesis. From (9), $\rho(\beta_0)$ can alternatively be written as a function of the true population ρ and the departure of the true β from the hypothesized value β_0, denoted $\Delta(\beta_0)$. Whereas Equation (6) is directly used for hypothesis testing via $\hat{\rho}(\beta_0)$, a consistent estimator of $\rho(\beta_0)$, the representation of $\rho(\beta_0)$ in (9) is used in the analysis of power below. When the null is true, $\beta = \beta_0$, $\Delta(\beta_0)$ equals zero; the mean of $t_{AR}(\beta_0)$ equals zero; and the correlation between $t_{AR}(\beta_0)$ and f simplifies to ρ.

Given the following algebraic expression for the square of the t-ratio,

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

the continuous mapping theorem yields

$$(10) \qquad \hat{t}^2 \xrightarrow{d} t^2 = t^2(t_{AR}(\beta_0), f, \rho(\beta_0)) \equiv \frac{t^2_{AR}(\beta_0)}{1 - 2\rho(\beta_0)\frac{t_{AR}(\beta_0)}{f} + \frac{t^2_{AR}(\beta_0)}{f^2}}.$$

This distribution will be important to the power analysis that follows.

IV.B Power curve analysis: VtF, tF, AR

We start with a standard power curve analysis comparing the performance of the three tests VtF, tF, and AR.

[32] Specifically, it is straightforward to derive that $\rho_{RF} = \frac{\frac{\sqrt{V(Zv)}}{\sqrt{V(Zu)}}\beta + \rho}{\sqrt{\left[\frac{V(Zv)}{V(Zu)}\beta^2 + 2\beta\frac{\sqrt{V(Zv)}}{\sqrt{V(Zu)}}\rho + 1\right]}}$ and $\sqrt{\frac{\sigma_{22}}{\sigma_{11}}} = \frac{\sqrt{V(Zv)}}{\sqrt{V(Zu)}}\sqrt{\frac{1}{\left[\frac{V(Zv)}{V(Zu)}\beta^2 + \sqrt{\frac{V(Zv)}{V(Zu)}}2\beta\rho + 1\right]}}$, and substitute these expressions into Equation (6).

The (asymptotic) critical region $\mathscr{R}(t_{AR}(\beta_0), f, \rho(\beta_0))$ for each of these tests is given by

$$\begin{aligned}
&\left\{t^2 > c(\rho(\beta_0), F; \alpha)\right\} \text{ for } VtF \\
&\left\{t_{AR}^2 > q_{1-\alpha}\right\} \text{ for } AR \\
&\left\{t^2 > c_{tF}(F; \alpha)\right\} \text{ for } tF,
\end{aligned}$$

where the critical region is expressed as depending on $t_{AR}(\beta_0)$, f, and $\rho(\beta_0)$ since t^2 is a function of these same arguments by (10). But, the joint distribution of $(t_{AR}(\beta_0), f)$ and the value of $\rho(\beta_0)$ in turn depend on the values of $\Delta(\beta_0)$, ρ, and f_0 as given in (9). Hence, the relevant data generating process for these tests can be indexed by $\Delta(\beta_0)$, ρ, and f_0. And, the power of each test can then be computed as the probability of rejection for each data generating process and hypothesized null value: $\Pr_{\Delta(\beta_0),\rho,f_0}[\mathscr{R}(t_{AR}(\beta_0), f, \rho(\beta_0))]$, which yields a power surface defined on a three-dimensional domain defined by the variables $\Delta(\beta_0)$, ρ, and f_0.

How one decides to "slice" this power surface into presentable two-dimensional curves has been carefully considered in previous work. Andrews, Marmer and Yu (2019) and Van de Sijpe and Windmeijer (2023) note that there are at least three different "slices" that could be considered. First, one could keep the reduced-form error covariance matrix and the null value β_0 constant while varying the true β (resulting in changes in the structural error covariance matrix[33] as true β varies over alternative values). Second, one could keep the structural error covariance matrix and the null β_0 constant and vary the true parameter β (causing the reduced-form error covariance matrix to vary with the alternative). A third possibility is to keep the true β, reduced-form and structural error covariance matrices constant while varying the hypothesized value β_0. As pointed out in LMMP (2022), in the context of the just-identified IV model, the second and third approaches produce identical power curves, and will differ from the curves using the first way of "slicing" the power surface.

While the interpretation and perspective provided by an individual power curve slice can vary from approach to approach, the collective information content of each approach is identical. That is, the collection of power curves of any of the three approaches described above simply summarizes the power surface given by $\Pr_{\Delta(\beta_0),\rho,f_0}[\mathscr{R}(t_{AR}(\beta_0), f, \rho(\beta_0))]$.

We adopt the second (equivalently, the third) approach to displaying power because these power curve slices are connected to the expected length of confidence sets and thus are the most closely associated with our results on confidence intervals discussed in previous sections. As shown by Pratt (1961), the integral of 1 minus the power (type II error), integrated across all values of β_0

[33] The structural errors referred to here are (u, v), see (1). The structural error covariance matrix can be obtained from the reduced form error matrix by pre- and post-multiplication by $\begin{pmatrix} 1 & -\beta \\ 0 & 1 \end{pmatrix}$ and its transpose.

while keeping β fixed, is equal to the expected length of the confidence set in repeated samples. The measure of length used is the usual Lebesgue measure. Pratt (1961) points out that expected length, which is easily recognizable from an applied perspective, has two interpretations: it represents both the average "size" of the set of false values that will be contained in the confidence set, as well as the average probability of the confidence set containing each false value (uniformly averaging across false values). By displaying power curves for each procedure across a wide range of β_0 values, we expect to gain insight into relative lengths, with the area between the curves specifically representing the expected difference in lengths. We also anticipate the curves to exhibit infinite areas between the power curve and the value 1. This is due to Dufour (1997) who shows that, in this weakly identified setting, any valid confidence set has infinite expected length. The correspondence of these areas above and between curves to expected length does not hold for the first approach (fixed reduced-form variance) to visualizing a "slice" of the power surface.

Figure 7: VtF Power Curves: $\rho = 0$, $f_0 = 1$

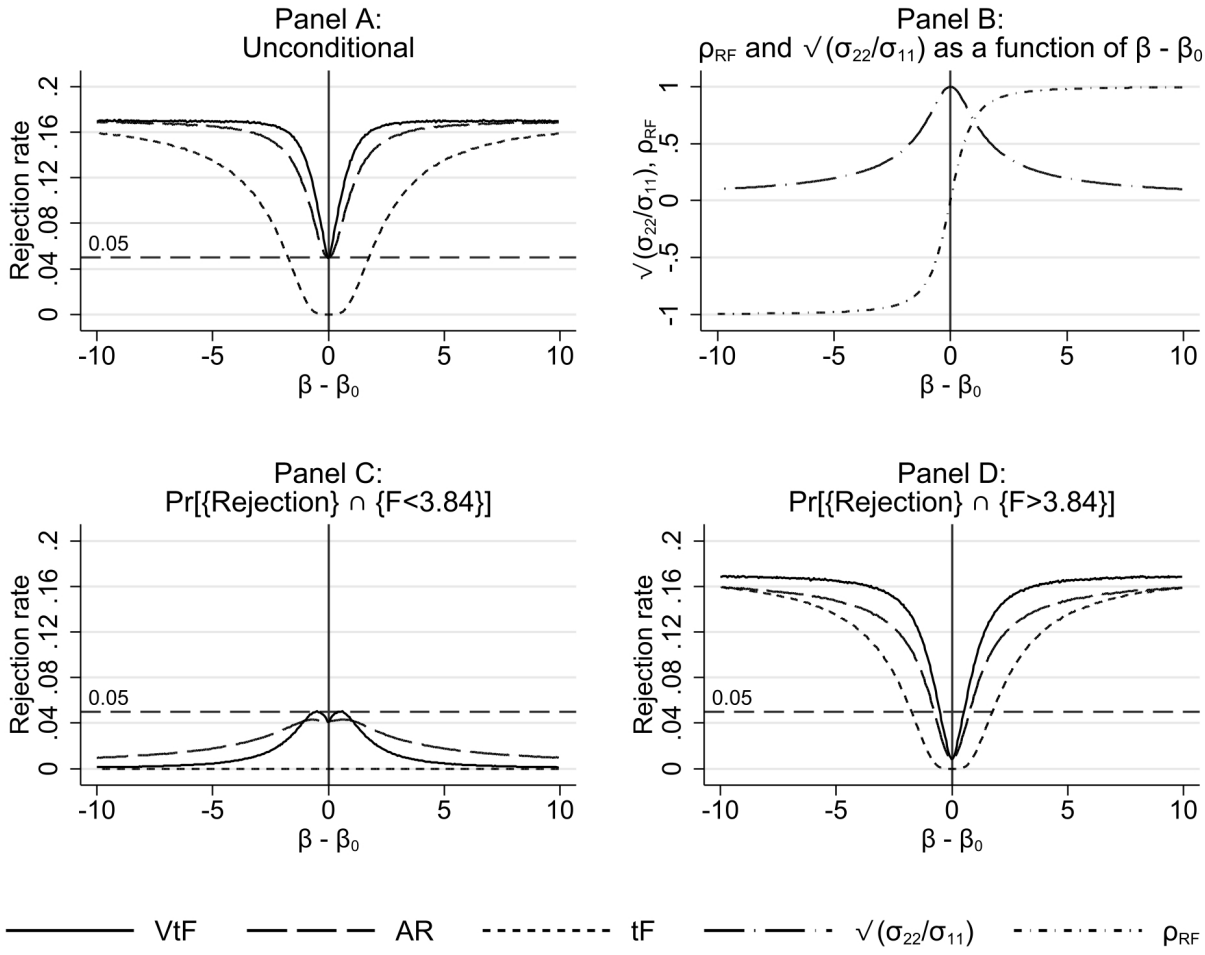
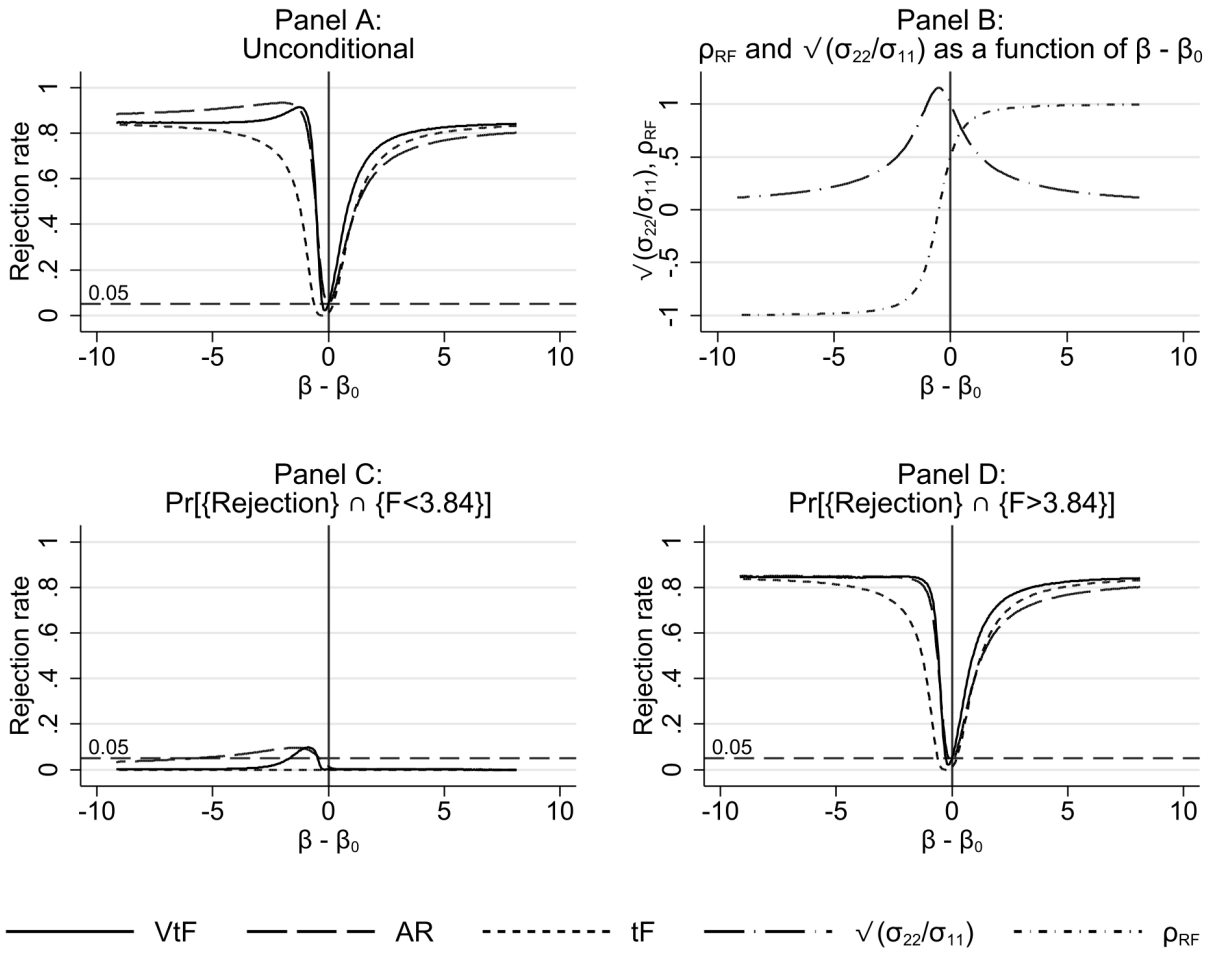


Figure 7 Panel A shows the second/third approach to visualizing power for VtF,tF, and AR

Figure 8: VtF Power Curves: $\rho = 0.5$, $f_0 = 3$



for the case where $\rho = 0$ and $f_0 = 1$. As a reminder that the reduced-form covariance matrix must necessarily change as $\beta - \beta_0$ varies, Panel B plots the corresponding values of $\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}$ and ρ_{RF}. Panel A clearly shows uniform dominance of VtF over tF across the range of parameter values $\beta - \beta_0$. tF is conservative in this case, as we would expect, since it must control rejection probabilities for the case of $\rho = \pm 1$ even if the true ρ is zero. Figure 7 Panel A also shows that VtF uniformly dominates AR over the same range.

Figure 8 Panel A displays the case of $\rho = 0.5$ and $f_0 = 3$. Once again, VtF uniformly dominates tF, but neither VtF nor AR uniformly dominate each other. The visual impression is that the area between the curves when VtF is higher is roughly similar to the area between the curves when AR is higher.

To consider how these power results fit with established AR optimality results, it is useful to first recall more precisely what these optimality results entail. For the just-identified IV model, AR

has been formally established as uniformly most powerful among unbiased tests (Moreira (2002, 2009)) and among invariant similar tests (Andrews, Moreira and Stock (2006)).[34]. From Figure 8 Panel A, we can see that neither tF nor VtF is unbiased. Though the bias of VtF is relatively small in this graph, it is sufficient to place VtF (and tF) outside the class of tests in which AR is uniformly most powerful. It is straightforward to show that VtF and tF also lie outside the invariant class of tests. So there is no contradiction between the power results of Figures 7 and 8 Panel A and the optimality of AR within unbiased or invariant similar classes of tests.

The uniformly most powerful invariant similar property implies admissibility of AR within the class of *all* valid tests (Andrews, Moreira and Stock (2006)), which *would* include VtF. In fact, the uniformly most powerful invariant similar property has been established pointwise in the reduced form error variance, so that admissibility of AR also holds for each fixed reduced form error variance and null. This means AR cannot be uniformly dominated along any power curves drawn according to the first approach which fixes the reduced form error variance and null. This result is entirely consistent with Figure 7 Panel A, which displays power via the second/third approach and not the first approach. For each value of $\beta - \beta_0$ there exists a completely different power curve that fixes the reduced form parameters $\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}$ and ρ_{RF} as given by Panel B. And on that particular slice (the first approach/slice, not pictured here), the admissibility result says that it can never be true that AR is uniformly dominated. Hence, the uniform dominance of VtF in Figure 7 Panel A is not at odds with the literature's existing admissibility result for AR.[35]

For completeness, in Appendix C.2, we provide the power curves for a range of different scenarios, covering $f_0 = 1, 3, 6, 9$, and $\rho = 0, .5, .8, .9$. In most cases, a visual inspection of the two curves suggests that VtF possesses higher power for some values of $\beta - \beta_0$ to about the same extent as AR has higher power for other values of $\beta - \beta_0$. However, the closer ρ is to zero, VtFs advantage over AR becomes more apparent.

IV.C Conditional power curve analysis: VtF, tF, AR

There is an important aspect of these traditional power curves that masks the relative performance of VtF and AR *confidence intervals.* The power curves show the probability of rejecting the hypothesis $\beta_0 \neq \beta$ regardless of whether $F > q_{1-\alpha}$. The condition $F > q_{1-\alpha}$ is crucial, because, the confidence sets for VtF, AR, and tF are bounded if and only if $\hat{F} > q_{1-\alpha}$, and our results on confidence interval performance, motivated by the notion that bounded intervals are the objects of

[34] These optimality results are generalized to the cases of heteroskedastic, clustered, and/or autocorrelated errors in Moreira and Moreira (2019)

[35] Moreira, Sharifvaghefi and Ridder (2021) show AR is optimal within a invariant similar class that allows the reduced form error covariance matrix to change and note a corresponding admissibility result. Not surprisingly, the VtF power results are also not contradictory to these findings.

interest to practitioners, focuses exclusively on the bounded confidence set region $\hat{F} > q_{1-\alpha}$.

To connect the results on power to the superior VtF confidence interval performance (shown in previous sections), it is illuminating to decompose the power curves into the asymptotic analogs of the unbounded and bounded confidence set conditions, $F \leq q_{1-\alpha}$ and $F > q_{1-\alpha}$:

$$\Pr_{\Delta(\beta_0),\rho,f_0}\left[\mathscr{R}\left(t_{AR}\left(\beta_0\right),f,\rho\left(\beta_0\right)\right)|F \leq q_{1-\alpha}\right]\cdot\Pr_{\Delta(\beta_0),\rho,f_0}\left[F \leq q_{1-\alpha}\right]$$

and

$$\Pr_{\Delta(\beta_0),\rho,f_0}\left[\mathscr{R}\left(t_{AR}\left(\beta_0\right),f,\rho\left(\beta_0\right)\right)|F > q_{1-\alpha}\right]\cdot\Pr_{\Delta(\beta_0),\rho,f_0}\left[F > q_{1-\alpha}\right]$$

where the two conditional power terms weighted by the unbounded and bounded confidence set probabilities sum to the unconditional power curves displayed in panel A of Figures 7 and 8. This allows us to decompose, for each alternative value of $\beta-\beta_0$, the extent to which the null is rejected when $F \leq q_{1-\alpha}$ (first-stage is statistically insignificant; confidence set unbounded) and when $F > q_{1-\alpha}$ (first-stage is statistically significant; confidence set bounded). In the weak instrument context, this decomposition is especially useful, because the region of unbounded confidence intervals is non-trivial and includes varying behavior among these methods.

The insight of Pratt (1961) can again be applied here so that

$$E\left[\text{length}_{\mathscr{R}}|F > q_{1-\alpha}\right] = \int\left(1-\Pr_{\Delta(\beta_0),\rho,f_0}\left[\mathscr{R}\left(t_{AR}\left(\beta_0\right),f,\rho\left(\beta_0\right)\right)|F > q_{1-\alpha}\right]\right)d\Delta\left(\beta_0\right)$$

where $\text{length}_{\mathscr{R}}$ is the length of the confidence set that corresponds to the decision rule to reject given by the test with critical region $\mathscr{R}$. The average length of the confidence set when it is bounded has the interpretation of the type II error – conditional on $F > q_{1-\alpha}$ – averaged over all false values β_0. It is worth noting that once we are comparing the conditional power of the various test procedures, we will be altogether removed from the original setting in which AR possessed the optimal power properties previously discussed.

Panels C and D of Figures 7 and 8 decompose the power curves of panel A. Panel C shows power for $F < q_{1-\alpha}$ corresponding to unbounded confidence sets. In this unbounded confidence set case, AR appears to have a significant advantage over VtF over a wide range of alternatives. Since the two components of conditional power must add to the unconditional power curves, we therefore would anticipate VtF to correspondingly have a more apparent advantage for the $F > q_{1-\alpha}$, and this is what is shown in panel D for both figures. In the case of $\rho = 0, f_0 = 1$, the uniform dominance of VtF over AR is even more pronounced when focusing on $F > q_{1-\alpha}$, the bounded confidence set region. In the case of $\rho = 0.5, f_0 = 3$, any power advantage of AR in the unconditional power curve (left side of panel A) essentially disappears, when focusing on just $F > q_{1-\alpha}$. For completeness, we show these graphs for other designs in Appendix C.2. Figure A2 reveals a

similar pattern: even though AR and VtF appear roughly balanced in terms of unconditional power across many designs, AR's power is more driven by rejecting when $F \leq q_{1-\alpha}$ (Panel B in Appendix C.2) and VtF's power is more driven by rejecting when $F > q_{1-\alpha}$ (Panel C in Appendix C.2).

The power curve comparisons with $F > q_{1-\alpha}$ correspond to our earlier results on confidence interval lengths. The power curves with $F \leq q_{1-\alpha}$ give insight into the power expended by these methods that goes toward forming unbounded confidence sets. For tF, unbounded confidence sets are always the entire real line, leading to zero power in this region. AR and VtF produce two types of unbounded confidence sets: the whole real line; and the real line excluding a bounded set of values; in the case of AR, it is a single bounded interval (often referred to as the "donut"). All of these unbounded confidence sets are two-sided, meaning that they will be incapable of rejecting the null hypotheses that $\beta_0 = \infty$ or $\beta_0 = -\infty$. While it may be possible that there are applied contexts where the ability to rule out intermediate values of the parameter while being unable to statistically rule out $\pm\infty$ (and generally extremely large positive and negative effects) is valuable, it is likely more common that researchers are interested in being able to statistically rule out large magnitudes in either direction, which is not possible with confidence sets that include $\pm\infty$.

We believe that it is informative to decompose power in this way to see whether the advantages of a given method derive from its unbounded or bounded confidence set behavior. Panel C of Figures 7 and 8 imply that AR's "donut holes" are generally larger than that of VtF, therefore leading to a power advantage in the *unbounded* confidence interval region. But AR's power advantage with unbounded confidence intervals comes at a power cost relative to VtF in the *bounded* confidence interval region. We find that VtF has a clear power advantage conditional on the bounded confidence region, which accords with our earlier findings from the data-realization by data-realization perspective.

For completeness, we also provide a comparison of distributions of confidence set lengths in Appendix C.1, see Table A7. The results shown there are not unexpected, given the findings in this section and section III.D. Table A7 produces quantiles of confidence interval length distributions across 16 designs for all of the methods considered here. The distribution of the differences in log lengths from Figures 6a-c are also shown here across these designs. All of these results reinforce VtF's confidence interval length advantage (see Appendix C.1 for additional details).

V Conclusion and Implications

The development of VtF is motivated by two aspects of the spirit of the work of Staiger and Stock (1997) and SY (2005), which has become the *de facto* industry standard for just-identified IV inference in the applied research community: 1) practitioners overwhelmingly use the t-ratio based on the 2SLS estimator along with a robust standard error, and would prefer to default to the

± 1.96 confidence intervals whenever possible; and 2) practitioners implicitly already use a data-dependent t-ratio critical value, by relying on a minimum threshold for the first stage F-statistic, which itself is an intuitive measure of instrument strength or weakness. VtF is designed on the basis of these two features, while also using the information provided by all the statistics of the model to try to obtain confidence intervals that are neither conservative nor anti-conservative, even when the first stage F-statistic is small.

After solving for the VtF critical value function, we assess whether the additional use of the statistic $\hat{r}$ leads to a meaningful gain in precision, by comparing its confidence interval performance to that of tF, which only uses $\hat{F}$. It does. As we comprehensively document in the heatmap in Figure 6a, and illustrate with our sample of empirical studies, VtF confidence intervals are considerably shorter than those of tF over a wide range of commonly-observed magnitudes of $\hat{F}$.

We recognize that our sample of specifications is an inherently selected one. For example, it is possible that the editorial process may be implicitly selecting on large F-statistics or statistically significant results. Furthermore, our sample represents only a small minority of IV studies: only 14 out of a possible 69 IV studies had reported the equivalent of the three regressions needed (2SLS, first-stage, reduced form) to compute $\hat{r}$. Therefore, it is an open question as to how results and conclusions might change for those remaining 55 out of 69 studies, if access to the micro-data were available. Assuming that $\hat{r}$ will not eventually be computed for those studies, we are unaware of any more powerful alternative to tF for drawing correct IV inferences when *only* $\hat{\beta}$, $\hat{se}(\hat{\beta})$ and the F-statistic is available.

However, moving forward, there is little reason to settle for tF when VtF is available. A clear implication of our findings for prevailing practice is that there are potentially substantial gains in precision that come along with reporting one additional and easy-to-compute statistic, $\hat{r}$.

Since VtF is entirely motivated from a practical perspective, we assessed the performance cost of tailoring a procedure to practitioner preferences; the findings are, in our view, unexpected. First, while seeking to characterize a region where the usual ± 1.96 confidence intervals can be validly used, we do find a substantial region of realizations of the data where VtF confidence sets are completely contained within the ± 1.96 confidence intervals. Since AR and tF confidence sets are always longer than ± 1.96 confidence intervals, this finding was at the outset unforeseen. As a result, one can use a simple rule of thumb $\hat{F} > 10 + 100 \cdot \hat{r}$ for using the usual ± 1.96 confidence intervals (and otherwise use VtF intervals), and maintain valid inference.

Second, when we compare the performance of VtF to the AR procedure, which is the standard recommendation from the econometric literature, we found that in all 89 specifications from our empirical studies, VtF produced shorter confidence intervals than AR. VtF's interval length advantage does not appear to be specific to our particular sample of specifications: this general advantage is also apparent in a comprehensive examination of all possible data realizations with

$1.96^2 < \hat{F} \leq 104.67$.

On their face, these findings appear to run counter to the conventional wisdom that AR is best for robust inference for just-identified IV. But we have been able to reconcile our findings with this conventional wisdom through two key points made in a power analysis of different procedures. The first point is that VtF just does not belong to the class of tests in which AR is uniformly most powerful. For instance, we show (Figure 8) that VtF is not an unbiased test. Unbiasedness of a test (i.e. the power curve never falling below the null rejection rate) is arguably desirable all other things equal. However, in this case, the requirement of unbiasedness of the test appears to be excluding procedures, that when inverted, deliver shorter confidence intervals.

We have also considered whether VtF's adherence to F-based t-ratio inference leads to a performance loss relative to the only other *biased* similar test for just-identified IV of which we are aware – the Conditional Wald test of Moreira (2003). In the Appendix, we show that there is no substantial performance loss of VtF, relative to Conditional Wald. Indeed Conditional Wald's power and confidence interval performance is quite similar to that of VtF; it, too, has similar power advantages over AR with generally shorter confidence sets. This constitutes further evidence that runs counter to the recommendation to use AR.[36] This additional finding on Conditional Wald suggests that if one seeks to find an even better-performing confidence interval procedure, by being willing to abandon the practitioner-friendly features of the F-based approach of SY (2005), as embodied in VtF, or abandon the t-ratio statistic entirely, then it seems prudent to include biased test procedures in any such search.

A second point is that confidence interval length comparisons in this weakly identified context are challenged by Dufour (1997)'s result that necessitates that all valid procedures produce unbounded confidence sets with positive probability. We speculate that applied researchers care very little about the distinction between confidence sets that are the whole real line and those that are unbounded but rule out a bounded set, since in both cases, either $+\infty$ or $-\infty$ cannot be ruled out. Moreover, typical unconditional power analyses essentially puts the size of the donut hole in an unbounded confidence set on equal footing with the length of a bounded confidence interval. To distinguish between these two cases, we introduce a corresponding decompositional analysis of *power* in the direction of bounded and unbounded confidence sets. These power decompositions provide an intuitive viewpoint on power curves for weakly identified sets and could be a useful consideration for attempts at identifying a procedure with optimal confidence set length properties.

[36] This finding on just-identified Conditional Wald is along the same lines as the that reported in Van de Sijpe and Windmeijer (2023), who present simulation evidence showing that, in the over-identified, homoskedastic case, Conditional Wald appears to outperform the Conditional Likelihood Ratio test of Moreira (2003).

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, Donald WK, Vadim Marmer, and Zhengfei Yu. 2019. "On optimal inference in the linear IV model." *Quantitative Economics*, 10: 457–485.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock. 2007. "Performance of Conditional Wald Tests in IV Regression with Weak Instruments." *Journal of Econometrics*, 139: 116–132.

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. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

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

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2003. "Instrumental variables and GMM: Estimation and testing." *The Stata Journal*, 3: 1–31.

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.

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

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." *arXiv preprint arXiv:2108.04887*.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and estimation of treatment effects with a regression-discontinuity design." *Econometrica*, 69: 201–209.

Hansen, Bruce. 2022. *Econometrics.* Princeton University Press.

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

Keane, Michael, and Timothy Neal. 2023. "Instrument strength in IV estimation and inference: A guide to theory and practice." *Journal of Econometrics*.

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

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

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter. 2022. "Valid t-ratio Inference for IV." *American Economic Review*, 112(10): 3260–3290.

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. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica*, 71: 1027–1048.

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

Moreira, Marcelo J, Mahrad Sharifvaghefi, and Geert Ridder. 2021. "Optimal invariant tests in an instrumental variables regression with heteroskedastic and autocorrelated errors." *arXiv preprint arXiv:1705.00231*.

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.

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

Stock, James, and Mark Watson. 2019. *Introduction to Econometrics (4th edition).* Addison Wesley Longman. Professor Stock receives royalties for this text.

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.

Van de Sijpe, Nicolas, and Frank Windmeijer. 2023. "On the power of the conditional likelihood ratio and related tests for weak-instrument robust inference." *Journal of Econometrics*, 1: 82–104.

Wooldridge, Jeffrey M. 2019. *Introductory econometrics: A modern approach.* Cengage learning.