ANNUAL REVIEWS

Annual Review of Economics

Weak Instruments in Instrumental Variables Regression: Theory and Practice

Isaiah Andrews,[1] James H. Stock,[1] and Liyang Sun[2]

[1]Department of Economics, Harvard University, Cambridge, Massachusetts 02138, USA; email: james_stock@harvard.edu

[2]Department of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Annu. Rev. Econ. 2019. 11:727–53

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-080218-025643

JEL codes: C18, C26

Keywords

weak instruments, heteroskedasticity, F-statistic

Abstract

When instruments are weakly correlated with endogenous regressors, conventional methods for instrumental variables (IV) estimation and inference become unreliable. A large literature in econometrics has developed procedures for detecting weak instruments and constructing robust confidence sets, but many of the results in this literature are limited to settings with independent and homoskedastic data, while data encountered in practice frequently violate these assumptions. We review the literature on weak instruments in linear IV regression with an emphasis on results for nonhomoskedastic (heteroskedastic, serially correlated, or clustered) data. To assess the practical importance of weak instruments, we also report tabulations and simulations based on a survey of papers published in the *American Economic Review* from 2014 to 2018 that use IV. These results suggest that weak instruments remain an important issue for empirical practice, and that there are simple steps that researchers can take to better handle weak instruments in applications.

1. INTRODUCTION

In instrumental variables (IV) regression, the instruments are called weak if their correlation with the endogenous regressors, conditional on any controls, is close to zero. When this correlation is sufficiently small, conventional approximations to the distribution of two-stage least squares and other IV estimators are generally unreliable. In particular, IV estimators can be badly biased, while t-tests may fail to control size, and conventional IV confidence intervals may cover the true parameter value far less often than intended.

Recognition of this problem has led to a great deal of work on econometric methods applicable to models with weak instruments. Much of this work, especially early in the literature, focused on the case where the data are independent, and the errors in the reduced-form and first-stage regressions are homoskedastic. Homoskedasticity implies that the variance matrix for the reduced-form and first-stage regression estimates can be written as a Kronecker product, which substantially simplifies the analysis of many procedures. As a result, there are now extensive theoretical results on detection of weak instruments and construction of identification-robust confidence sets in the homoskedastic case.

More recently, much of the theoretical literature on weak instruments has considered the more difficult case where the data may be dependent and/or the errors heteroskedastic. In this setting, which we refer to as the nonhomoskedastic case, the variance of the reduced-form and first-stage estimates no longer has Kronecker product structure in general, rendering results based on such structure inapplicable. Because homoskedasticity is rarely a plausible assumption in practice, procedures applicable to the nonhomoskedastic case have substantial practical value.

This review focuses on the effects of weak instruments in the nonhomoskedastic case. We concentrate on detection of weak instruments and weak-instrument-robust inference. The problem of detection is relevant because weak-instrument-robust methods can be more complicated to use than standard two-stage least squares, so if instruments are plausibly strong, then it is convenient to report two-stage least squares estimates and standard errors. If instruments are weak, however, then practitioners are advised to use weak-instrument-robust methods for inference, the second topic of this review. We do not survey estimation, an area in which less theoretical progress has been made.[1]

In addition to surveying the theoretical econometrics literature, we examine the role of weak instruments in empirical practice using a sample of 230 specifications gathered from 17 papers published in the *American Economic Review* (AER) from 2014 to 2018 that use the word "instrument" in their abstract. These papers use a wide variety of instruments to study a broad range of questions. For example, Hornung (2014) studies the long-term effects of skilled migration using the settlement patterns of French Huguenots in Prussia, instrumenting with population losses due to plagues during the Thirty Years' War. Young (2014) studies the effect of sectoral growth on total factor productivity, instrumenting with defense expenditures. Favara & Imbs (2015) study the effect of bank credit on housing prices, instrumenting with US bank branching deregulations. A full list of the papers that we consider, as well as additional details, are contained in the **Supplemental Appendix**.

Supplemental Material >

We use this sample for two purposes. The first is to learn what empirical researchers are actually doing when it comes to detecting and handling weak instruments. The second is to develop a

[1] Two notable exceptions are the work of Hirano & Porter (2015) and Andrews & Armstrong (2017). Hirano & Porter's (2015) contribution is a negative result: They prove that, if one includes the possibility that instruments can be arbitrarily weak, then no unbiased estimator of the coefficient of interest exists without further restrictions. Andrews & Armstrong (2017) show, however, that if one imposes correctly the sign of the first-stage regression coefficient, then asymptotically unbiased estimation is possible, and they derive unbiased estimators.

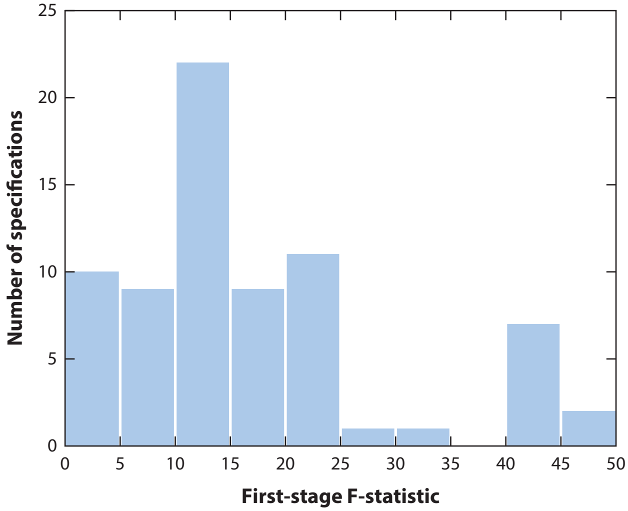

Figure 1

Distribution of reported first-stage F-statistics (and their nonhomoskedastic generalizations) in 72 specifications with a single endogenous regressor and first-stage F smaller than 50. The total number of single endogenous regressor specifications reporting F-statistics is 108.

collection of specifications that we use to assess the importance of weak instrument issues and the performance of weak instrument methods in data-generating processes reflective of real-world settings.

Figure 1 displays a histogram of first-stage F-statistics reported in specifications in our AER sample with a single endogenous regressor, truncated at 50 for visibility. The first-stage F-statistic for testing the hypothesis that the instruments are unrelated to the endogenous regressor is a standard measure of the strength of the instrument. Many of the first-stage F-statistics in our AER sample are in a range that, based on simulations and theoretical results, raises concerns about weak instruments, including many values less than 10. This suggests that weak instruments are frequently encountered in practice.

Another noteworthy feature of the data underlying **Figure 1** is that 13 of the 16 papers in our sample with a single endogenous regressor reported at least one first-stage F-statistic. Evidently, and reassuringly, there is widespread recognition by researchers that one needs to be attentive to the potential problems caused by weak instruments. This said, our review of these papers leads us to conclude that there is room for improving current empirical practice.

Specifically, in the leading case with a single endogenous regressor, we recommend that researchers judge instrument strength based on the effective F-statistic of Montiel Olea & Pflueger (2013). If there is only a single instrument, then we recommend reporting identification-robust Anderson-Rubin (AR) confidence intervals. These are efficient regardless of the strength of the instruments, and so should be reported regardless of the value of the first-stage F. Finally, the literature has not yet converged on a single procedure for dealing with multiple instruments (i.e., in the overidentified case) but we recommend choosing from among the several available robust procedures that are efficient when the instruments are strong.

The review is organized as follows. Section 2 lays out the IV model and notation. Section 3 describes the weak instruments problem. Section 4 reviews methods for detecting weak instruments, Section 5 reviews weak-instrument-robust inference, and Section 6 concludes with a discussion of

Supplemental Material >

open questions in the literature on weak instruments. In the **Supplemental Appendix**, we discuss our AER sample, the details of our simulation designs, and available Stata implementations of the procedures that we discuss in the main text.

2. THE INSTRUMENTAL VARIABLES MODEL

We study the linear IV model with a scalar outcome Y_i, a $p \times 1$ vector of potentially endogenous regressors X_i, a $k \times 1$ vector of instrumental variables Z_i, and an $r \times 1$ vector of controls W_i. This yields the linear constant effects IV model

$$Y_i = X_i'\beta + W_i'\kappa + \varepsilon_i, \tag{1.}$$

$$X_i' = Z_i'\pi + W_i'\gamma + V_i', \tag{2.}$$

where Equation 1 is the structural equation, while Equation 2 is a linear projection commonly called the first stage. We are interested in β, but X_i is potentially endogenous, so regression of Y_i on X_i and W_i may yield a biased estimate of β. We assume that Z_i is a valid instrument after controlling for W_i, and in particular, that for $Z_i^{\perp}$, the residual from projecting Z_i on W_i, $E[Z_i^{\perp}\varepsilon_i] = 0$. Furthermore, we have $E[Z_iV_i'] = 0$ and $E[W_iV_i'] = 0$ by the definition of linear projection.

We allow the possibility that the errors (ε_i, V_i) are conditionally heteroskedastic given the exogenous variables (Z_i, W_i), so $E[(\varepsilon_i, V_i')'(\varepsilon_i, V_i')|Z_i, W_i]$ may depend on (Z_i, W_i). We further allow the possibility that (Y_i, X_i, Z_i, W_i) are dependent across i, for example, due to clustering or time-series correlation. Finally, the results that we discuss generalize to the case where the data are nonidentically distributed across i, although for simplicity, we do not pursue this extension.

Substituting for X_i in Equation 1, we obtain the equation

$$Y_i = Z_i'\delta + W_i'\tau + U_i \tag{3.}$$

with $\delta = \pi\beta$. In a common abuse of terminology, we refer to Equation 1 as the structural form, Equation 2 as the first stage, and Equation 3 as the reduced form (for the older meaning of these terms, see, e.g., Hausman 1983). We can equivalently express the model as Equations 1 and 2 or as Equations 2 and 3, since each set of equations is an invertible linear transformation of the other. Likewise, the errors $(U_i, V_i) = (\varepsilon_i + \beta' V_i, V_i)$ are an invertible linear transformation of (ε_i, V_i).

For ease of exposition, we focus primarily on the case with a scalar endogenous regressor X_i, and so assume $p = 1$ unless noted otherwise. In our AER sample, 211 of the 230 specifications have $p = 1$, so this appears to be the leading case in practice. Furthermore, unless noted otherwise, we assume that the instruments Z_i are orthogonal to the control variables W_i, and so drop the controls from our notation. We discuss how to handle nonorthogonal control variables at the end of this section.

In this review, we focus on estimators and tests that are functions of the reduced-form least squares coefficient $\hat{\delta}$, the first-stage least squares coefficient $\hat{\pi}$, and matrices that can be consistently estimated from the first stage and reduced form (e.g., variance and weighting matrices). Estimators in this class include two-stage least squares, which for $\hat{Q}_{ZZ} = \frac{1}{n}\sum Z_iZ_i'$ can be written as

$$\hat{\beta}_{2\text{SLS}} = \left(\hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}\right)^{-1}\hat{\pi}'\hat{Q}_{ZZ}\hat{\delta}, \tag{4.}$$

as well as efficient-two-step generalized method of moments (GMM) $\hat{\beta}_{2\text{SGMM}} = [\hat{\pi}'\hat{\Omega}(\hat{\beta}^1)^{-1}\hat{\pi}]^{-1}\hat{\pi}'\hat{\Omega}(\hat{\beta}^1)^{-1}\hat{\delta}$, with $\hat{\Omega}(\beta)$ as an estimator for the variance of $\hat{\delta} - \hat{\pi}\beta$ and $\hat{\beta}^1$ as a first-step estimator. Limited information maximum likelihood and continuously updated GMM likewise fall into this class.

Under mild regularity conditions (and, in the time-series case, stationarity), $(\hat{\delta}, \hat{\pi})$ are consistent and asymptotically normal in the sense that

$$\sqrt{n}\begin{pmatrix} \hat{\delta}-\delta \\ \hat{\pi}-\pi \end{pmatrix} \rightarrow_d N\left(0, \Sigma^*\right) \qquad 5.$$

for

$$\Sigma^* = \begin{pmatrix} \Sigma^*_{\delta\delta} & \Sigma^*_{\delta\pi} \\ \Sigma^*_{\pi\delta} & \Sigma^*_{\pi\pi} \end{pmatrix} = \begin{pmatrix} Q^{-1}_{ZZ} & 0 \\ 0 & Q^{-1}_{ZZ} \end{pmatrix} \Lambda^* \begin{pmatrix} Q^{-1}_{ZZ} & 0 \\ 0 & Q^{-1}_{ZZ} \end{pmatrix},$$

where $Q_{ZZ} = E[Z_i Z_i']$ and

$$\Lambda^* = \lim_{n\to\infty} Var\left[\left(\frac{1}{\sqrt{n}}\sum_i U_i Z_i', \frac{1}{\sqrt{n}}\sum_i V_i Z_i'\right)'\right].$$

Thus, the asymptotic variance of $\sqrt{n}(\hat{\delta}-\delta, \hat{\pi}-\pi)$ has the usual sandwich form. Under standard assumptions, the sample-analog estimator $\hat{Q}_{ZZ}$ will be consistent for Q_{ZZ}, and we can construct consistent estimators $\hat{\Lambda}^*$ for Λ^*. These results imply the usual asymptotic properties for IV estimators. For example, assuming that the constant-effect IV model is correctly specified (so $\delta = \pi\beta$), and that π is fixed and nonzero, the delta method, together with Equation 5, implies that $\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \rightarrow_d N(0, \Sigma^*_{\beta,2SLS})$ for $\Sigma^*_{\beta,2SLS}$ consistently estimable. We can likewise use Equation 5 to derive the asymptotic distribution for other IV estimators, such as limited information maximum likelihood or two-step and continuously updated GMM.

2.1. Homoskedastic and Nonhomoskedastic Cases

A central distinction in the literature on weak instruments, and in the historical literature on IV more broadly, is between what we term the homoskedastic and nonhomoskedastic cases. In the homoskedastic case, we assume that the data (Y_i, X_i, Z_i, W_i) are independent and identically distributed (IID) across i, and the errors (U_i, V_i) are homoskedastic, so $E[(U_i, V_i)'(U_i, V_i)|Z_i, W_i]$ does not depend on (Z_i, W_i). Whenever these conditions fail, whether due to heteroskedasticity or dependence (e.g., clustering or time-series dependence), we say we are in the nonhomoskedastic case.

Two-stage least squares is efficient in the homoskedastic case but not, in general, in the nonhomoskedastic case. Whether homoskedasticity holds also determines the structure of Λ^*. Specifically, in the homoskedastic case, we can write

$$\Lambda^* = E\left[\begin{pmatrix} U_i^2 & U_i V_i \\ U_i V_i & V_i^2 \end{pmatrix} \otimes (Z_i Z_i')\right] = E\left[\begin{pmatrix} U_i^2 & U_i V_i \\ U_i V_i & V_i^2 \end{pmatrix}\right] \otimes Q_{ZZ},$$

where the first equality follows from the assumption of IID data, while the second follows from homoskedasticity. Thus, in homoskedastic settings, the variance matrix Λ^* can be written as the Kronecker product of a 2 × 2 matrix that depends on the errors with a $k \times k$ matrix that depends on the instruments. The matrix Σ^* inherits the same structure, which, as we note below, simplifies several calculations. By contrast, in the nonhomoskedastic case, Σ^* does not in general have Kronecker product structure, rendering these simplifications inapplicable.

2.2. Dealing with Control Variables

If the controls W_i are not orthogonal to the instruments Z_i, then we need to take them into account. In this more general case, let us define $(\hat{\delta}, \hat{\pi})$ as the coefficients on Z_i from the reduced-form and first-stage regressions of Y_i and X_i, respectively, on (Z_i, W_i). By the Frisch-Waugh theorem, these are the same as the coefficients from regressing Y_i and X_i on $Z_i^{\perp}$ (again the part of Z_i orthogonal to W_i). One can likewise derive estimators for the asymptotic variance matrix Σ^* in terms of $Z_i^{\perp}$ and suitably defined regression residuals. Such estimators, however, necessarily depend on the assumptions imposed on the data-generating process (for example, whether we allow heteroskedasticity, clustering, or time-series dependence).

A simple way to estimate Σ^* in practice when there are control variables is to jointly estimate $(\hat{\delta}, \hat{\pi})$ in a seemingly unrelated regression with whatever specification one would otherwise use (including fixed effects and clustering or serial-correlation robust standard errors). Appropriate estimates of Σ^* are then generated automatically by standard statistical software.

3. THE WEAK INSTRUMENTS PROBLEM

Motivated by the asymptotic approximation in Equation 5, let us consider the case where the reduced-form and first-stage regression coefficients are jointly normal,

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta \\ \pi \end{pmatrix}, \Sigma\right), \qquad 6.$$

with $\Sigma = \frac{1}{n}\Sigma^*$ known (and, for ease of exposition, full rank). Effectively, Equation 6 discards the approximation error in Equation 5, as well as the estimation error in $\hat{\Sigma}^*$, to obtain a finite-sample normal model with known variance. This suppresses any complications arising from nonnormality of the ordinary least squares (OLS) estimates or difficulties with estimating Σ and focuses attention solely on the weak instruments problem. Correspondingly, results derived in the model in Equation 6 will provide a good approximation to behavior in applications where the normal approximation to the distribution of $(\hat{\delta}, \hat{\pi})$ is accurate and Σ is well-estimated. By contrast, in settings where the normal approximation is problematic or $\hat{\Sigma}$ is a poor estimate of Σ, results derived based on Equation 6 will be less reliable (see Section 6; see also Young 2018).

Since the IV model implies that $\delta = \pi\beta$, the IV coefficient is simply the constant of proportionality between the reduced-form coefficient δ and the first-stage parameter π. In the just-identified setting, matters simplify further, with the IV coefficient becoming $\beta = \delta/\pi$ and the usual IV estimators, including two-stage least squares and GMM, simplifying to $\hat{\beta} = \hat{\delta}/\hat{\pi}$. Just-identified specifications with a single endogenous variable constitute a substantial fraction of the specifications in our AER sample (101 out of 230), highlighting the importance of this case in practice.

It has long been understood (see, e.g., Fieller 1954) that ratio estimators like $\hat{\beta}$ can behave badly when the denominator is close to zero. The weak instruments problem is simply the generalization of this issue to potentially multidimensional settings. In particular, when the first-stage coefficient π is close to zero relative to the sampling variability of $\hat{\pi}$, the normal approximations to the distribution of IV estimates discussed in Section 2 may be quite poor. Nelson & Startz (1990a,b) provided early simulation demonstrations of this issue, while Bound et al. (1995) found similar issues in simulations based on the work of Angrist & Krueger (1991).

The usual normal approximation to the distribution of $\hat{\beta}$ can be derived using the delta method, which linearizes $\hat{\beta}$ in $(\hat{\delta}, \hat{\pi})$. Under this linear approximation, normality of $(\hat{\delta}, \hat{\pi})$ implies approximate normality of $\hat{\beta}$. This normal approximation fails in settings with weak instruments because $\hat{\beta}$ is highly nonlinear in $\hat{\pi}$ when the latter is close to zero. As a result, normality of $(\hat{\delta}, \hat{\pi})$ does not imply approximate normality of $\hat{\beta}$. Specifically, the IV coefficient $\hat{\beta} = \hat{\delta}/\hat{\pi}$ is distributed as the ratio of potentially correlated normals, and so is nonnormal. If π is large relative to the standard error of $\hat{\pi}$, however, then $\hat{\pi}$ falls close to zero with only very low probability, and the nonlinearity of $\hat{\beta}$ in $(\hat{\delta}, \hat{\pi})$ ceases to matter. Thus, we see that nonnormality of the IV estimator arises when the first-stage parameter π is small relative to its sampling variability. The same issue arises in the overidentified case with $p = 1 < k$, where the weak instruments problem arises when the $k \times 1$ vector π of first-stage coefficients is close to zero relative to the variance of $\hat{\pi}$. Likewise, in the general $1 \leq p \leq k$ case, the weak instruments problem arises when the $k \times p$ matrix π of first-stage coefficients is close to having reduced rank relative to the sampling variability of $\hat{\pi}$.

3.1. Failure of the Bootstrap

A natural suggestion for settings where conventional asymptotic approximations fail is the bootstrap. Unfortunately, the bootstrap (and its generalizations, including subsampling and the m-out-of-n bootstrap) does not in general resolve weak instruments issues (see Andrews & Guggenberger 2009). For intuition, note that we can view the bootstrap as simulating data based on estimates of the data-generating process. In the model in Equation 6, the worst case for identification is $\pi = 0$, since in this case, β is totally unidentified. We never estimate π perfectly, however, and in particular, we estimate $\hat{\pi} = 0$ with probability zero. Thus, the bootstrap incorrectly thinks that β is identified with probability one. More broadly, the bootstrap can make systematic errors in estimating the strength of the instruments, which suggests the reason why it can yield unreliable results. None of the IV specifications in our AER sample used the bootstrap.

3.2. Motivation of the Normal Model

The normal model in Equation 6 has multiple antecedents. Several papers in the early econometric literature on simultaneous equations assumed fixed instruments and exogenous variables along with normal errors, which leads to the homoskedastic version of Equation 6, sometimes with Σ unknown (Anderson & Rubin 1949, Mariano & Sawa 1972, Sawa 1969).

More recently, several papers in the literature on weak instruments, including those of Kleibergen (2002), Moreira (2003), Andrews et al. (2006), and Moreira & Moreira (2015), derive results in the normal model in Equation 6, sometimes with the additional assumption that the underlying data are normal. While in this case we have motivated the normal model in Equation 6 heuristically based on the asymptotic normality (Equation 5) of the reduced-form and first-stage estimates, this connection is made precise elsewhere in the literature. Staiger & Stock (1997) show that the normal model in Equation 6 arises as an approximation to the distribution of the scaled reduced-form and first-stage regression coefficients under weak-instrument asymptotics, where the first-stage shrinks at a $\sqrt{n}$ rate. As discussed by Staiger & Stock (1997), these asymptotics are intended to capture situations in which the true value of the first stage is on the same order as sampling uncertainty in $\hat{\pi}$, so issues associated with small π cannot be ignored. Finite sample results for the model in Equation 6 then translate to weak-instrument asymptotic results via the continuous mapping theorem. Many other authors, including Kleibergen (2005), Andrews et al. (2006), Andrews (2016), and Andrews & Armstrong (2017), have built on these results to prove validity for particular procedures under weak-instrument asymptotics.

More recently, Andrews & Guggenberger (2015, 2017), Andrews & Mikusheva (2016), Andrews (2017), and Andrews et al. (2018) have considered asymptotic validity uniformly over values of the first-stage parameter π and distributions for (U_i, V_i, W_i, Z_i). These authors show that some, although not all, procedures derived in the normal model in Equation 6 are also uniformly asymptotically valid in the sense that, e.g., the probability of incorrectly rejecting true null hypotheses converges to the nominal size uniformly over a large class of data-generating processes as the sample size increases. Andrews et al. (2018) discuss general techniques to establish uniform asymptotic validity, but the argument for a given procedure is case specific. Thus, in this review, we focus on the normal model in Equation 6, which unites much of the weak-instruments literature, and refer readers interested in questions of uniformity to the papers cited above.

3.3. Simulated Distribution of t-Statistics

While we know from theory that weak instruments can invalidate conventional inference procedures, whether weak instruments are a problem in a given application is necessarily case specific. To examine the practical importance of weak instruments in recent applications of IV methods, we report simulation results calibrated to our AER sample.

Specifically, we calibrate the normal model in Equation 6 to the 124 specifications in our AER sample for which we estimate a positive-definite variance matrix Σ of the reduced-form and first-stage estimates based either on results reported in the paper or on replication files. This excludes specifications without replication data where we cannot estimate Σ from results reported in the text, as well as four specifications where our estimate of Σ is not positive definite.[2] It happens to be the case that all remaining specifications have only a single endogenous regressor ($p = 1$). Thus, our simulation results only address this case. In each specification, we set the first-stage parameter π to the estimate $\hat{\pi}$ in the data and set δ to $\hat{\pi}\hat{\beta}_{2SLS}$, the product of the first stage with the two-stage least squares estimates. We set Σ equal to the estimated variance matrix for $(\hat{\delta}, \hat{\pi})$, maintaining whatever assumptions were used by the original authors (including the same controls and clustering at the same level).

In each specification, we repeatedly draw first-stage and reduced-form parameter estimates $(\hat{\delta}^*, \hat{\pi}^*)$ and, for each draw, calculate the two-stage least squares estimate, along with the t-statistic for testing the true value of β (that is, the value used to simulate the data). In **Figures 2*a*** and **3*a***, we plot the median t-statistic and the frequency with which nominal 5% two-sided t-tests reject on the vertical axis, and the average of the effective F-statistic of Montiel Olea & Pflueger (2013), which we introduce in the next section, on the horizontal axis. This statistic is equivalent to the conventional first-stage F-statistic for testing $\pi = 0$ in models with homoskedastic errors but adds a multiplicative correction in models with nonhomoskedastic errors. For visibility, we limit attention to the 106 out of 124 specifications where the average first-stage F-statistic is smaller than 50 (the remaining specifications exhibit behavior very close to those with F-statistics between 40 and 50).

Several points emerge clearly from these results. First, there is a nontrivial number of specifications with small first-stage F-statistics [e.g., below 10, the rule-of-thumb cutoff for weak instruments proposed by Staiger & Stock (1997)] in the AER data. Second, even for specifications with essentially the same first-stage F-statistic, the median t-statistic and the size of nominal 5% t-tests can vary substantially due to other features (for example, the true value β and the matrix Σ). Third, we see that, among specifications with a small average F-statistic, behavior can deviate substantially from what we would predict under conventional (strong-instrument) asymptotic approximations. Specifically, conventional approximations imply that the median t-statistic is zero, and 5% t-tests should reject 5% of the time. In our simulations, by contrast, we see that the median

[2] Three appear due to rounding error in cases where we calculate Σ based on reported estimates and standard errors, while the last arises from a case with a large number of fixed effects and a small number of clusters.

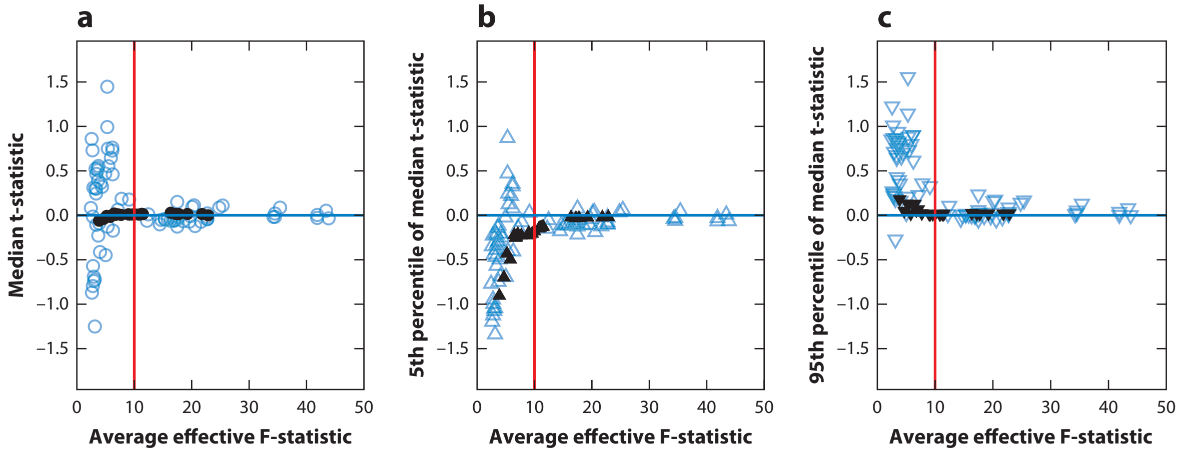

Figure 2

Median of t-statistic for testing true value of β plotted against the average effective F-statistic of Montiel Olea & Pflueger (2013) in calibrations to *American Economic Review* (AER) data, limited to the 106 out of 124 specifications with average F smaller than 50. Just-identified specifications are plotted as solid black, while overidentified specifications are plotted as blue outlines. (*a*) Median at parameter values estimated from AER data. (*b*) 5th and (*c*) 95th percentiles of the median t-statistic under the Bayesian exercise described in the text. The vertical red line in each panel corresponds to a first-stage F of 10.

t-statistic sometimes has absolute value larger than one, while the size of 5% t-tests can exceed 30%. These issues largely disappear among specifications where the average F-statistic exceeds 10, and in these cases, conventional approximations appear to be fairly accurate.

These results suggest that weak-instrument issues are relevant for modern applications of IV methods. It is worth emphasizing that these simulations are based on the normal model in

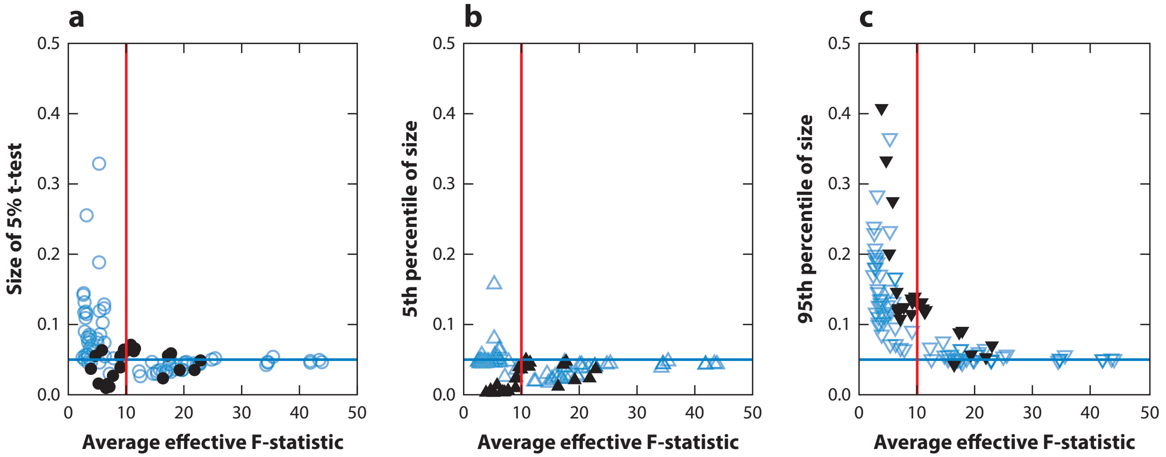

Figure 3

Rejection probability for nominal 5% two-sided t-tests plotted against the average effective F-statistic of Montiel Olea & Pflueger (2013) in calibrations to *American Economic Review* (AER) data, limited to the 106 out of 124 specifications with average F smaller than 50. Just-identified specifications are plotted as solid black, while overidentified specifications are plotted as blue outlines. (*a*) Size at parameter values estimated from AER data. (*b*) 5th and (*c*) 95th percentiles of size under the Bayesian exercise described in the text. The vertical red line in each panel corresponds to a first-stage F of 10.

Equation 6 with known variance Σ, so these results arise from the weak instruments problem alone and not from, e.g., nonnormality of $(\hat{\delta}, \hat{\pi})$ or difficulties estimating the variance matrix Σ.

These results are sensitive to the parameter values considered (indeed, this is the reason that the bootstrap fails). Since we estimate (β, π) with error, it is useful to quantify the uncertainty around our estimates for the median t-statistic and the size of t-tests. To do so, we adopt a Bayesian approach consistent with the normal model in Equation 6 and simulate a posterior distribution for the median t-statistic and the size of 5% t-tests. Specifically, we calculate the posterior distribution on (δ, π) after observing $(\hat{\delta}, \hat{\pi})$ using the normal likelihood from Equation 6 and a flat prior. We draw values

$$\begin{pmatrix} \tilde{\delta} \\ \tilde{\pi} \end{pmatrix} \sim N\left(\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix}, \Sigma\right)$$

for the reduced-form and first-stage parameters from this posterior, calculate the implied two-stage least squares coefficient $\tilde{\beta}$, and repeat our simulations taking $(\tilde{\beta}, \tilde{\pi})$ to be the true parameter values (setting the reduced-form coefficient to $\tilde{\pi}\tilde{\beta}$). **Figures 2*b*** and **3*b*** report the 5th percentiles of the median t-statistic and size, respectively, across draws $(\tilde{\beta}, \tilde{\pi})$, while **Figures 2*c*** and **3*c*** report the 95th percentiles. As these results suggest, there is considerable uncertainty about the distribution of t-statistics in these applications. As in our baseline simulations, however, poor performance for conventional approximations is largely, although not exclusively, limited to specifications where the average effective F-statistic is smaller than 10.

Finally, it is interesting to consider behavior when we limit attention to the subset of specifications that are just identified (i.e., that have $k = 1$), which are plotted as solid black in **Figures 2** and **3**. Interestingly, when we simulate behavior at parameter estimates from the AER data in these cases, we find that the largest absolute median t-statistic is 0.06, while the maximal size for a 5% t-test is just 7.1%. If, however, we consider the bounds from our Bayesian approach, the worst-case absolute median t-statistic is 0.9, while the worst-case size for the t-test is over 40%. Thus, t-statistics appear to behave much better in just-identified specifications when we consider simulations based on the estimated parameters, but this is no longer the case once we incorporate uncertainty about the parameter values.

4. DETECTING WEAK INSTRUMENTS

The simulation results in Section 3 suggest that weak instruments may render conventional estimates and tests unreliable in a nontrivial fraction of published specifications. This raises the question of how to detect weak instruments in applications. A natural initial suggestion is to test the hypothesis that the first stage is equal to zero, $\pi = 0$. As noted by Stock & Yogo (2005), however, conventional methods for inference on β are unreliable not only for $\pi = 0$, but also for π in a neighborhood of zero. Thus, we may reject that $\pi = 0$ even when conventional inference procedures are unreliable. To overcome this issue, we need formal procedures for detecting weak instruments, rather than tests for total nonidentification.

4.1. Tests for Weak Instruments with Homoskedastic Errors

Stock & Yogo (2005) consider the problem of testing for weak instruments in cases with homoskedastic errors. They begin by formally defining the set of values π that they call weak. They consider two different definitions, the first based on the bias of IV estimates relative to OLS and the second based on the size of Wald- or t-tests. In each case, they include a value of π in the weak instrument set if the worst-case bias or size over all possible values of β exceeds a threshold (they phrase this result in terms of the correlation between the errors ε and V in Equations 1 and 2, but for Σ known, this is equivalent). They then develop formal tests for the null hypothesis that the instruments are weak (that is, that π lies in the weak instrument set), where rejection allows one to conclude that the instruments are strong.

In settings with a single endogenous regressor, Stock & Yogo's (2005) tests are based on the first-stage F-statistic. Their critical values for this statistic depend on the number of instruments, and tables are available in the work of Stock & Yogo (2005). If we define the instruments as weak when the worst-case bias of two-stage least squares exceeds 10% of the worst-case bias of OLS, then the results of Stock & Yogo show that, for between 3 and 30 instruments, the appropriate critical value for a 5% test of the null of weak instruments ranges from 9 to 11.52 and so is always close to the Staiger & Stock (1997) rule-of-thumb cutoff of 10. By contrast, if we define the instruments as weak when the worst-case size of a nominal 5% t-test based on two-stage least squares exceeds 15%, then the critical value depends strongly on the number of instruments and is equal to 8.96 in cases with a single instrument but rises to 44.78 in cases with 30 instruments.

Stock & Yogo (2005) also consider settings with multiple endogenous variables. For such cases, they develop critical values for use with the Cragg & Donald (1993) statistic for testing the hypothesis that π has reduced rank. Building on these results, Sanderson & Windmeijer (2016) consider tests for whether the instruments are weak for the purposes of estimation and inference on one of multiple endogenous variables.

4.2. Tests for Weak Instruments with Nonhomoskedastic Errors

The results of Stock & Yogo (2005) rely heavily on the assumption of homoskedasticity. As discussed above, in homoskedastic settings, the variance matrix Σ for $(\hat\delta, \hat\pi)$ can be written as the Kronecker product of a 2×2 matrix with a $k \times k$ matrix, which Stock & Yogo (2005) use to obtain their results. As noted in Section 2, by contrast, Σ does not in general have Kronecker product structure in nonhomoskedastic settings, and the tests of Stock & Yogo (2005) do not apply. Specifically, in the nonhomoskedastic case, the homoskedastic first-stage F-statistic is inapplicable and should not be compared to the Stock & Yogo (2005) critical values (Montiel Olea & Pflueger 2013).

Despite the inapplicability of Stock & Yogo's (2005) results, F-statistics are frequently reported in nonhomoskedastic settings with multiple instruments. In such cases, some authors report nonhomoskedasticity-robust F-statistics

$$F_{\mathrm{R}} = \frac{1}{k}\hat\pi' \hat\Sigma_{\pi\pi}^{-1}\hat\pi, \qquad 7.$$

for $\hat\Sigma_{\pi\pi}$ an estimator for the variance of $\hat\pi$, while others report traditional, nonrobust F-statistics

$$F_{\mathrm{N}} = \frac{1}{k}\hat\pi' \hat\Sigma_{\pi\pi,N}^{-1}\hat\pi = \frac{n}{k\hat\sigma_V^2}\hat\pi' \hat Q_{ZZ}\hat\pi \qquad 8.$$

for $\hat\Sigma_{\pi\pi,N} = \frac{\hat\sigma_V^2}{n}\hat Q_{ZZ}^{-1}$ and $\hat\sigma_V^2$ an estimator for $E[V_i^2]$. In our AER data, for instance, none of the 52 specifications that both have multiple instruments and report first-stage F-statistics assume homoskedasticity to calculate standard errors for $\hat\beta$, but at least six report F-statistics that do assume homoskedasticity (we are unable to determine the exact count because most authors do not explicitly describe how they calculate F-statistics, and not all papers provide replication data). To illustrate, **Figure 4*a*** plots the distribution of F-statistics reported in papers in our AER sample, broken down by the method (robust or nonrobust) used, when we can determine this. Given the mix of methods, we use F-statistic as a generic term to refer both to formal first-stage F-statistics

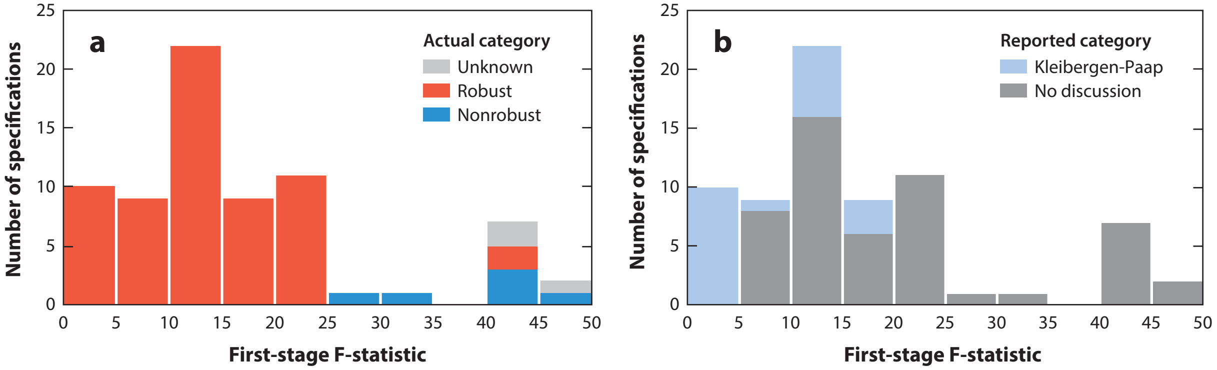

Figure 4

Distribution of reported first-stage F-statistics (and their nonhomoskedastic generalizations) in 72 specifications with a single endogenous regressor and first-stage F smaller than 50. Thirty-six other specifications (not shown) have a single endogenous regressor but first-stage F-statistic larger than 50. (*a*) Decomposition by statistic computed (either nonrobust F-statistic F_N, robust F-statistic F_R, or unknown). Note that, in settings with a single endogenous regressor, the Kleibergen-Paap F-statistic reduces to the robust F-statistic, so we categorize papers reporting this statistic accordingly. (*b*) Decomposition by label used by authors in text (either Kleibergen-Paap or not explicitly discussed).

F_N (which assume homoskedasticity and single endogenous regressor) and to generalizations of F-statistics to nonhomoskedastic settings, cases with multiple endogenous regressors, and so on.

Use of F-statistics in nonhomoskedastic settings is built into common statistical software. When run without assuming homoskedastic errors, the popular ivreg2 command in Stata automatically reports the Kleibergen & Paap (2007) Wald statistic for testing that π has reduced rank along with critical values based on Stock & Yogo (2005) (Baum et al. 2007), although the output warns users about Stock & Yogo's (2005) homoskedasticity assumption. In settings with a single endogenous variable, the Kleibergen & Paap (2007) Wald statistic is equivalent to a nonhomoskedasticity-robust F-statistic F_R for testing $\pi = 0$, while in settings with multiple endogenous regressors, it is a robust analog of the Cragg & Donald (1993) statistic. Interestingly, despite the equivalence of Kleibergen-Paap statistics and robust F-statistics in settings with a single endogenous variable, the distribution of published F-statistics appears to differ depending on what label the authors use. In particular, as shown in **Figure 4*b***, published F-statistics labeled by authors as Kleibergen-Paap statistics tend to be smaller.

We are unaware of theoretical justification for the use of either F_N or F_R to gauge instrument strength in nonhomoskedastic settings. As an alternative, Montiel Olea & Pflueger (2013) propose a test for weak instruments based on the effective first-stage F-statistic

$$F_{\mathrm{Eff}} = \frac{\hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}}{tr(\hat{\Sigma}_{\pi\pi}\hat{Q}_{ZZ})} = \frac{k\hat{\sigma}_V^2/n}{tr(\hat{\Sigma}_{\pi\pi}\hat{Q}_{ZZ})}F_N = \frac{tr(\hat{\Sigma}_{\pi\pi,N}\hat{Q}_{ZZ})}{tr(\hat{\Sigma}_{\pi\pi}\hat{Q}_{ZZ})}F_N. \qquad 9.$$

In cases with homoskedastic errors, F_{Eff} reduces to F_N, while in cases with nonhomoskedastic errors it incorporates a multiplicative correction that depends on the robust variance estimate. Likewise, in the just-identified case, F_{Eff} reduces the F_R [and so also coincides with the Kleibergen & Paap (2007) Wald statistic], while in the nonhomoskedastic case, it weights $\hat{\pi}$ by $\hat{Q}_{ZZ}$ rather than $\hat{\Sigma}_{\pi\pi}^{-1}$.

The expressions for the two-stage least squares estimator in Equation 4, F_R in Equation 7, F_N in Equation 8, and F_{Eff} in Equation 9 provide some intuition for why F_{Eff} is an appropriate statistic for testing instrument strength when using two-stage least squares in the nonhomoskedastic case, while F_{R} and F_{N} are not. Two-stage least squares behaves badly when its denominator, $\hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}$, is close to zero. The statistic F_{N} measures this same object, but, because it is nonrobust, it gets the standard error wrong and so does not have a noncentral χ^2 distribution, as in the work of Stock & Yogo (2005). Indeed, in the nonhomoskedastic case, F_{N} can be extremely large with high probability even when $\pi' Q_{ZZ}\pi$ is small. By contrast, the statistic F_{R} measures the wrong population object, $\pi'\Sigma_{\pi\pi}^{-1}\pi$ rather than $\pi' Q_{ZZ}\pi$, so while it has a noncentral χ^2 distribution, its noncentrality parameter does not correspond to the distribution of $\hat{\beta}_{2\mathrm{SLS}}$.[3] Finally, F_{Eff} measures the right object and gets the standard errors right on average. More precisely, F_{Eff} is distributed as a weighted average of noncentral χ^2 variables where the weights, given by the eigenvalues of $\hat{\Sigma}_{\pi\pi}^{\frac{1}{2}}\hat{Q}_{ZZ}\hat{\Sigma}_{\pi\pi}^{\frac{1}{2}}/tr(\hat{\Sigma}_{\pi\pi}\hat{Q}_{ZZ})$, are positive and sum to one. Montiel Olea & Pflueger (2013) show that the distribution of F_{Eff} can be approximated by a noncentral χ^2 distribution and formulate tests for weak instruments as defined based on the Nagar (1959) approximation to the bias of two-stage least squares and limited information maximum likelihood. Their test rejects when the effective F-statistic exceeds a critical value. Note, however, that their argument is specific to two-stage least squares and limited information maximum likelihood, so if one were to use a different estimator, a different test would be needed.

For $k = 1$, $\Sigma_{\pi\pi}$, $\Sigma_{\pi\pi,N}$, and Q_{ZZ} are all scalar, and $F_{\mathrm{R}} = F_{\mathrm{Eff}}$. Both statistics have a noncentral χ^2 distribution with the same noncentrality parameter that governs the distribution of the IV estimator. Thus, in settings with $k = 1$, $F_{\mathrm{R}} = F_{\mathrm{Eff}}$ can be used with the Stock & Yogo (2005) critical values based on t-test size (the mean of the IV estimate does not exist when $k = 1$).

For $k > 1$, as noted above, the theoretical results of Montiel Olea & Pflueger (2013) formally concern only the Nagar (1959) approximation to the bias. Our simulations based on the AER data reported in Section 3 suggest, however, that effective F-statistics may convey useful information about instrument strength more broadly, since we see that conventional asymptotic approximations appear reasonable in specifications where the average effective F-statistic exceeds 10. This is solely an empirical observation about a particular data set, but why this is the case in these data and whether this finding generalizes to a broader range of empirically relevant settings are interesting questions for future research.

The main conclusion from this section is that F_{Eff}, not F_{R} or F_{N}, is the preferred statistic for detecting weak instruments in overidentified, nonhomoskedastic settings with one endogenous variable where one uses two-stage least squares or limited information maximum likelihood.[4] F_{Eff} can be compared to Stock & Yogo (2005) critical values for $k = 1$ and to Montiel Olea & Pflueger (2013) critical values for $k > 1$, or to the rule-of-thumb value of 10. It appears that none of the papers in our AER sample computed F_{Eff} (except for the $k = 1$ case where it equals F_{R}), but we hope to see wider use of this statistic in the future.

[3] The inapplicability of F_{R} and F_{N} in the nonhomoskedastic case is illustrated by the following example, which builds on an example of Montiel Olea & Pflueger (2013). Let $k = 2$, $Q_{ZZ} = I_2$, and $\Sigma_{\pi\pi} = E\left[\begin{pmatrix} U_i^2 & U_iV_i \\ U_iV_i & V_i^2 \end{pmatrix}\right] \otimes \begin{pmatrix} \omega^2 & 0 \\ 0 & \omega^{-2} \end{pmatrix}$. Under weak-instrument asymptotics with $\pi = C/\sqrt{n}$ for C fixed with both elements nonzero, as $\omega^2 \to \infty$, one can show that the distribution of the two-stage least squares estimate is centered around the probability limit of OLS, which is what we expect in the fully unidentified case. Thus, from the perspective of two-stage least squares, the instruments are irrelevant asymptotically. At the same time, both F_{N} and F_{R} diverge to infinity and so will indicate that the instruments are strong with probability one. By contrast, F_{Eff} converges to a χ_1^2 and so correctly reflects that the instruments are weak for the purposes of two-stage least squares estimation.

[4] Unfortunately, we are unaware of an analog of the Montiel Olea & Pflueger (2013) approach for settings with multiple endogenous variables.

4.3. Screening on the First-Stage F-Statistic

Given a method for detecting weak instruments, there is a question of what to do if we decide that the instruments are weak. Anecdotal evidence and our AER data suggest that, in some instances, researchers or journals may decide that specifications with small first-stage F-statistics should not be published. Specifically, **Figure 1** shows many specifications just above the Staiger & Stock (1997) rule-of-thumb cutoff of 10, consistent with selection favoring F-statistics above this threshold.

It is important to note that **Figure 1** limits attention to specifications where the original authors report first-stage F-statistics and uses the F-statistics as reported by the authors. By contrast, in our simulation results, we calculate effective F-statistics for all specifications in our simulation sample (i.e., where we can obtain a positive definite estimate of the variance matrix Σ), including in specifications where the authors do not report F-statistics, and match the assumptions used to calculate F-statistics to those used to calculate standard errors on $\hat{\beta}$. So, for example, in a paper that assumed homoskedastic errors to calculate F-statistics, but nonhomoskedastic errors to calculate standard errors on $\hat{\beta}$, we use a nonhomoskedasticity-robust estimator $\hat{\Sigma}_{\pi\pi}$ to compute the effective F-statistic in our simulations but report the homoskedastic F-statistic F_{N} in **Figure 1**. We do this because the F-statistic reported by the original authors seems to be the relevant one when thinking about selection on F-statistics.

While selection on first-stage F-statistics is intuitively reasonable, it can unfortunately result in bias in published estimates and size distortions in published tests. This point was made early in the weak instruments literature by Zivot et al. (1998) and relates to issues of pretesting and publication bias more generally. To illustrate the impact of these issues, we consider simulations calibrated to our AER data in which we drop all simulation draws where the effective F-statistic is smaller than 10. **Figure 5** plots the size of nominal 5% t-tests in this setting against the average effective F-statistic (where the average effective F-statistic is calculated over all simulation draws, not just those with $F_{\mathrm{Eff}} > 10$).

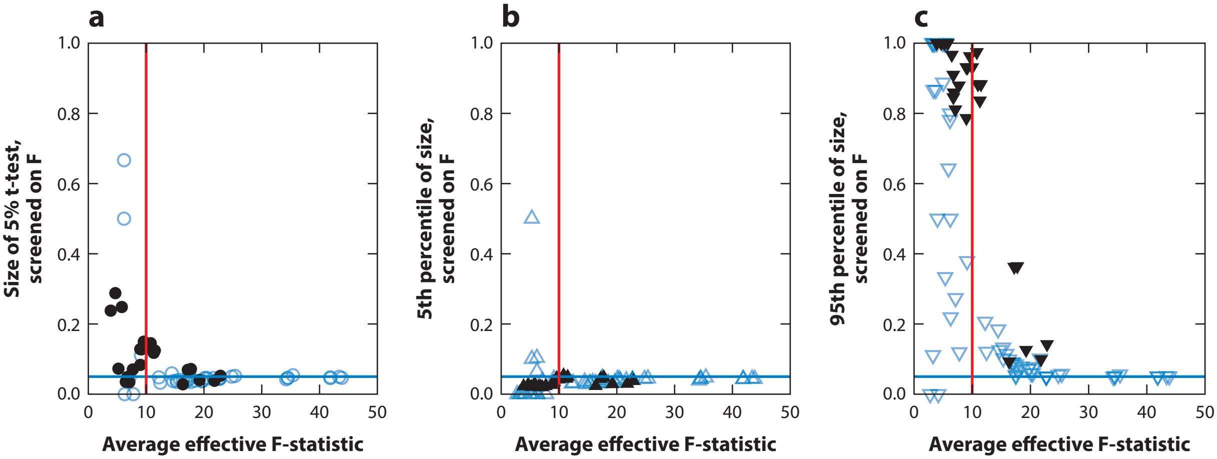

Figure 5

Rejection probability for nominal 5% two-sided t-tests after screening on $F_{\mathrm{Eff}} > 10$, plotted against the average effective F-statistic in calibrations to *American Economic Review* (AER) data, limited to the 106 out of 124 specifications with average effective F smaller than 50. Just-identified specifications are plotted as solid black, while overidentified specifications are plotted as blue outlines. (*a*) Size at parameter values estimated from AER data. (*b*) 5th and (*c*) 95th percentiles of size under the Bayesian exercise described in Section 3. The vertical red line in each panel corresponds to a first-stage F of 10.

The results in **Figure 5** highlight that screening on the F-statistics can dramatically increase size distortions. This is apparent even in simulations based on reported parameter estimates (shown in **Figure 5*a***), where the maximal size exceeds 70%, as compared to a maximal size of less than 35% for t-tests without screening on the F-statistic. Matters look still worse when considering the upper bound for size (shown in **Figure 5*c***), where many specifications have size close to one. Thus, screening on the first-stage F-statistic appears to compound, rather than reduce, inferential problems arising from weak instruments. This problem is not specific to the effective F-statistic F_{Eff} and also appears if we screen on F_{N} or F_{R}. Likewise, if we move the threshold from 10 to some other value, then we continue to see size distortions in a neighborhood of the new threshold.

If we are confident that our instruments are valid but are concerned they may be weak, then screening on F-statistics is unappealing for another reason: It unnecessarily eliminates specifications of potential economic interest. In particular, as we discuss in the next section, a variety of procedures for identification-robust inference on β have been developed in the literature. By using these procedures, we may gain insight from the data even in settings where the instruments are weak. Thus, weak instruments alone are not a reason to discard applications.

5. INFERENCE WITH WEAK INSTRUMENTS

The literature on weak instruments has developed a variety of tests and confidence sets that remain valid whether or not the instruments are weak, in the sense that their probability of incorrectly rejecting the null hypothesis and covering the true parameter value remains well-controlled. Since IV estimates are nonnormally distributed when the instruments are weak, these procedures do not rely on point estimates and standard errors but instead use test inversion.

The idea of test inversion is that, if we are able to construct a size-α test of the hypothesis $H_0 : \beta = \beta_0$ for any value β_0, then we can construct a level $1 - \alpha$ confidence set for β by collecting the set of nonrejected values. Formally, let us represent a generic test of $H_0 : \beta = \beta_0$ by $\phi(\beta_0)$, where we write $\phi(\beta_0) = 1$ if the test rejects and $\phi(\beta_0) = 0$ otherwise. We say that $\phi(\beta_0)$ is a size-α test of $H_0 : \beta = \beta_0$ in the normal model in Equation 6 if

$$\sup_{\pi} E_{\beta_0,\pi}\left[\phi(\beta_0)\right] \leq \alpha,$$

so the maximal probability of rejecting the null hypothesis, assuming that the null is true, is bounded above by α no matter the value of π. If $\phi(\beta_0)$ is a size-α test of $H_0 : \beta = \beta_0$ for all values β_0, then $CS = \{\beta : \phi(\beta) = 0\}$, the set of values not rejected by ϕ, is a level $1 - \alpha$ confidence set

$$\inf_{\beta,\pi} Pr_{\beta,\pi}\left\{\beta \in CS\right\} \geq 1 - \alpha. \qquad 10.$$

In practice, we can implement test inversion by taking a grid of potential values β, evaluating the test ϕ at all values in the grid, and approximating our confidence set by the set of nonrejected values.

When the instruments can be arbitrarily weak, correct coverage in Equation 10 turns out to be a demanding requirement. Specifically, the results of Gleser & Hwang (1987) and Dufour (1997) imply that in the normal model in Equation 6 without restrictions on (β, π), any level $1 - \alpha$ confidence set for β must have infinite length with positive probability. For intuition, consider the case in which $\pi = 0$, so β is unidentified. In this case, the data are entirely uninformative about β, and to ensure coverage $1 - \alpha$, a confidence set CS must cover each point in the parameter space with at least this probability, which is impossible if CS is always bounded. That the confidence set must be infinite with positive probability for all (β, π) then follows from the fact that the normal distribution has full support. Thus, if the event $\{CS$ infinite length$\}$ has positive probability under $\pi = 0$, then the same is true under all (β, π). This immediately confirms that we cannot obtain correct coverage under weak instruments by adjusting our (finite) standard errors, and so points to the need for a different approach such as test inversion.

To fix ideas, we first discuss test inversion based on the AR statistic, which turns out be efficient in just-identified models with a single instrument. We then turn to alternative procedures developed for overidentified models and inference on subsets of parameters. Finally, we discuss the effect of choosing between robust and nonrobust procedures based on a pretest for instrument strength. Since we base our discussion on the OLS estimates $(\hat{\delta}, \hat{\pi})$, the procedures that we discuss can be viewed as minimum-distance identification-robust procedures, as in the work of Magnusson (2010).

5.1. Inference for Just-Identified Models: The Anderson-Rubin Test

Test inversion offers a route forward in models with weak instruments because the IV model with parameter β implies restrictions on the distribution of the data regardless of the strength of the instruments. Specifically, the IV model implies that $\delta = \pi\beta$. Thus, under a given null hypothesis $H_0 : \beta = \beta_0$, we know that $\delta - \pi\beta_0 = 0$ and thus that

$$g(\beta_0) = \hat{\delta} - \hat{\pi}\beta_0 \sim N(0, \Omega(\beta_0)) \text{ for } \Omega(\beta_0) = \Sigma_{\delta\delta} - \beta(\Sigma_{\delta\pi} + \Sigma_{\pi\delta}) + \beta^2\Sigma_{\pi\pi},$$

where $\Sigma_{\delta\delta}$, $\Sigma_{\pi\pi}$, and $\Sigma_{\delta\pi}$ denote the variance of $\hat{\delta}$, the variance of $\hat{\pi}$, and their covariance, respectively. Thus, the AR statistic (Anderson & Rubin 1949), defined as $AR(\beta) = g(\beta)'\Omega(\beta)^{-1}g(\beta)$, follows a χ^2_k distribution under $H_0 : \beta = \beta_0$ no matter the value of π. Note that Anderson & Rubin (1949) considered the case with homoskedastic normal errors, so the AR statistic as we define it in this review is formally a generalization of their statistic that allows for nonhomoskedastic errors.

Using the AR statistic, we can form an AR test of $H_0 : \beta = \beta_0$ as $\phi_{AR}(\beta_0) = 1\{AR(\beta_0) > \chi^2_{k,1-\alpha}\}$ for $\chi^2_{k,1-\alpha}$, the $1 - \alpha$ quantile of a χ^2_k distribution. As noted by Staiger & Stock (1997), this yields a size-α test that is robust to weak instruments. Thus, if we were to recompute **Figure 3** for the AR test, then the size would be flat at 5% for all specifications. We can thus form a level $1 - \alpha$ weak-instrument-robust confidence set CS_{AR} by collecting the nonrejected values. In the case with homoskedastic errors (or with nonhomoskedastic errors but a single instrument), as noted by, e.g., Mikusheva (2010), one can derive the bounds of CS_{AR} analytically, avoiding the need for numerical test inversion.

Since AR confidence sets have correct coverage regardless of the strength of the instruments, we know from Gleser & Hwang (1987) and Dufour (1997) that they have infinite length with positive probability. Specifically, as discussed by Dufour & Taamouti (2005) and Mikusheva (2010), CS_{AR} can take one of three forms in settings with a single instrument: (*a*) a bounded interval $[a, b]$, (*b*) the real line $(-\infty, \infty)$, or (*c*) the real line excluding an interval $(-\infty, a] \cup [b, \infty)$. In settings with more than one instrument but homoskedastic errors, the AR confidence set can take the same three forms or may be empty. These behaviors are counterintuitive but have simple explanations.

First, as noted by Kleibergen (2007), as $|\beta| \to \infty$, $AR(\beta)$ converges to the Wald statistic for testing that $\pi = 0$ (equal to k times the robust first-stage F-statistic). Thus, the level-α AR confidence set has infinite length if and only if a robust F-test cannot reject that $\pi = 0$, and thus that β is totally unidentified. Thus, infinite-length confidence sets arise exactly in those cases where the data do not allow us to conclude that β is identified at all.

Second, CS_{AR} may be empty only in the overidentified setting. In this case, the AR approach tests that $\delta = \pi\beta_0$, which could fail either because $\delta = \pi\beta$ for $\beta \neq \beta_0$ or because there exists no value β such that $\delta = \pi\beta$. In the latter case, the overidentifying restrictions of the IV model fail. Thus, the AR test has power against both violations of our parametric hypothesis of interest and violations of the IV model's overidentifying restrictions, and an empty AR confidence set can be interpreted as a rejection of the overidentifying restrictions. The overidentifying restrictions could fail due either to invalidity of the instruments or to treatment effect heterogeneity, as in the work of Imbens & Angrist (1994), but in either scenario, the constant-effect IV model is misspecified.

The power of AR tests against violations of the IV model's overidentifying restrictions means that, if we care only about power for testing the parametric restriction $H_0 : \beta = \beta_0$, then AR tests and confidence sets can be inefficient. In particular, in the strongly identified case with $\|\pi\|$ large, one can show that the usual Wald statistic $(\hat{\beta} - \beta_0)^2/\hat{\sigma}^2_{\hat{\beta}}$ is approximately noncentral-χ^2_1 distributed with the same noncentrality as $AR(\beta_0)$, so tests based on the Wald statistic (or, equivalently, two-sided t-tests) have higher power than tests based on AR. Strong identification is important for this result. Chernozhukov et al. (2009) show that the AR test is admissible (i.e., not dominated by any other test) in settings with homoskedastic errors and weak instruments.

5.1.1. Efficiency of Anderson-Rubin in just-identified models. In just-identified models, there are no overidentifying restrictions, and the AR test has power only against violations of the parametric hypothesis. In this setting, Moreira (2009) shows that the AR test is uniformly most powerful unbiased. We say that a size-α test ϕ is unbiased if $E_{\beta,\pi}[\phi(\beta_0)] \geq \alpha$ for all $\beta \neq \beta_0$ and all π, so that the rejection probability when the null hypothesis is violated is at least as high as the rejection probability when the null is correct. The AR test is unbiased, and Moreira (2009) shows that, for any other size-α unbiased test ϕ, $E_{\beta,\pi}[\phi_{\text{AR}}(\beta_0) - \phi(\beta_0)] \geq 0$ for all $\beta \neq \beta_0$ and all π. Thus, the AR test has (weakly) higher power than any other size-α unbiased test no matter the true value of the parameters. In the strongly identified case, the AR test is asymptotically efficient in the usual sense and so does not sacrifice power relative to the conventional t-test.

5.1.2. Practical performance of Anderson-Rubin confidence sets. Since AR confidence sets are robust to weak identification and are efficient in the just-identified case, there is a strong case for using these procedures in just-identified settings. To examine the practical impact of using AR confidence sets, we return to our AER data set, limiting attention to just-identified specifications with a single endogenous variable where we can estimate the joint variance–covariance matrix of $(\hat{\pi}, \hat{\delta})$. In the sample of 34 specifications meeting these requirements, we find that AR confidence sets are quite similar to t-statistic confidence sets in some cases but are longer in others. Specifically, in two specifications, the first stage is not distinguishable from zero at the 5% level, so AR confidence sets are infinite. In the remaining 32 specifications, AR confidence sets are 56.5% longer than t-statistic confidence sets on average, although this difference drops to 20.3% if we limit attention to specifications that report a first-stage F-statistic larger than 10 and to 0.04% if we limit attention to specifications that report a first-stage F-statistic larger than 50. Complete results are reported in the **Supplemental Appendix**, Section D.

Supplemental Material >

5.2. Tests for Overidentified Models

In contrast to the just-identified case, in overidentified settings, the AR test is robust but inefficient under strong identification. This has led to a large literature seeking procedures that perform better in overidentified models.

Toward this end, note that, in the normal model in Equation 6, the AR statistic for testing $H_0 : \beta = \beta_0$ depends on the data only through $g(\beta_0) = \hat{\delta} - \hat{\pi}\beta_0$. To construct procedures that perform as well as the t-test in the strongly identified case, it is valuable to incorporate information from

$\hat{\pi}$, which is informative about which deviations of $\delta - \pi\beta_0$ from zero correspond to violations of the parametric restrictions of the model, rather than the overidentifying restrictions. Specifically, under alternative parameter value β, we have $\hat{\delta} - \hat{\pi}\beta_0 \sim N(\pi(\beta - \beta_0), \Omega(\beta_0))$ (for discussion, see Andrews 2016). Thus, to construct procedures that perform as well as the t-test in well-identified, overidentified cases, several authors have considered test statistics that depend on $(\hat{\delta}, \hat{\pi})$ through more than $\hat{\delta} - \hat{\pi}\beta_0$.

Once we seek to construct weak-instrument-robust tests that depend on the data through more than $g(\beta_0)$, however, we encounter an immediate problem: Even under the null $H_0 : \beta = \beta_0$, the distribution of $(\hat{\delta}, \hat{\pi})$ depends on the (unknown) first-stage parameter π. Thus, for a generic test statistic $s(\beta_0)$ that depends on $(\hat{\delta}, \hat{\pi})$, the distribution of $s(\beta_0)$ under the null will typically depend on π. For example, if we take $s(\beta_0)$ to be the absolute t-statistic $|\hat{\beta} - \beta_0|/\hat{\sigma}_{\hat{\beta}}$, then we know that the distribution of t-statistics under the null depends on the strength of the instruments. One could in principle find the largest possible $1 - \alpha$ quantile for $s(\beta_0)$ over the null consistent with some set of values for π, for example, an initial confidence set, as in the Bonferroni approach of Staiger & Stock (1997). For many statistics $s(\beta_0)$, however, this requires extensive simulation and will be computationally intractable; moreover, it typically entails a loss of power.

An alternative approach eliminates dependence on π through conditioning. Specifically, under $H_0 : \beta = \beta_0$, we have

$$\begin{pmatrix} g(\beta_0) \\ \hat{\pi} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \pi \end{pmatrix}, \begin{pmatrix} \Omega\left(\beta_0\right) & \Sigma_{\delta\pi} - \Sigma_{\pi\pi}\beta_0 \\ \Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta_0 & \Sigma_{\pi\pi} \end{pmatrix} \right).$$

Thus, if we define

$$D(\beta) = \hat{\pi} - (\Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta)\,\Omega(\beta)^{-1} g(\beta),$$

then we see that $(g(\beta), D(\beta))$ is a one-to-one transformation of $(\hat{\delta}, \hat{\pi})$, and under $H_0 : \beta = \beta_0$,

$$\begin{pmatrix} g(\beta_0) \\ D(\beta_0) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \pi \end{pmatrix}, \begin{pmatrix} \Omega\left(\beta_0\right) & 0 \\ 0 & \Psi(\beta_0) \end{pmatrix} \right)$$

for $\Psi(\beta_0) = \Sigma_{\pi\pi} - (\Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta_0)\,\Omega\left(\beta_0\right)^{-1} (\Sigma_{\delta\pi} - \Sigma_{\pi\pi}\beta_0)$. Thus, under the null, the nuisance parameter π enters the distribution of the data only through the statistic $D(\beta_0)$, while $g(\beta_0)$ is independent of $D(\beta_0)$ and has a known distribution. Thus, the conditional distribution of $g(\beta_0)$ [and thus of $(\hat{\delta}, \hat{\pi})$] given $D(\beta_0)$ does not depend on π. This conditioning approach was initially introduced to the weak instruments literature by Moreira (2003), who studied the homoskedastic case. In settings with homoskedastic errors, $g(\beta_0)$ and $D(\beta_0)$ are transformations of the statistics S and T introduced by Moreira (2003) (see Andrews & Mikusheva 2016).

We can simulate the conditional distribution of any statistic $s(\beta_0)$ given $D(\beta_0)$ under the null by drawing $g(\beta_0)^* \sim N(0, \Omega(\beta_0))$, constructing $(\hat{\delta}^*, \hat{\pi}^*)$ as

$$\begin{pmatrix} \hat{\delta}^* \\ \hat{\pi}^* \end{pmatrix} = \begin{pmatrix} I + \beta_0\left(\Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta_0\right)\Omega\left(\beta_0\right)^{-1} & \beta_0 I \\ \left(\Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta_0\right)\Omega\left(\beta_0\right)^{-1} & I \end{pmatrix} \begin{pmatrix} g\left(\beta_0\right)^* \\ D\left(\beta_0\right) \end{pmatrix}$$

for given $D\left(\beta_0\right)$, and tabulating the resulting distribution of $s^*(\beta_0)$ calculated based on $(\hat{\delta}^*, \hat{\pi}^*)$. If we denote the conditional $1 - \alpha$ quantile as $c_\alpha(D(\beta_0))$, we can then construct a conditional test based on s as $\phi_s = 1\{s(\beta_0) > c_\alpha(D(\beta_0))\}$, and provided that $s(\beta_0)$ is continuously distributed conditional on $D(\beta_0)$, this test has rejection probability exactly α under the null, $E_{\beta_0,\pi}\left[\phi_s(\beta_0)\right] = \alpha$ for all π; if the conditional distribution of $s(\beta_0)$ has point masses, then the test has size less than or equal to α. As noted by Moreira (2003) for the homoskedastic case, this allows us to construct a size-α

test based on any test statistic s. For further discussion of the simulation approach described above and a formal size control result applicable to the nonhomoskedastic case, the reader is referred to Andrews & Mikusheva (2016).

Tests that have rejection probability exactly α for all parameter values consistent with the null are said to be similar. Lehmann & Romano (2005, theorem 4.3) imply that, if the set of values of π is unrestricted, then all similar size-α tests of $H_0 : \beta = \beta_0$ are conditional tests in the sense that their conditional rejection probability given $D(\beta_0)$ under the null is equal to α. Moreover, in the present setting, the power functions for all tests are continuous, so if the set of values (β, π) is unrestricted, then all unbiased tests are necessarily similar. Thus, the class of conditional tests nests the class of unbiased tests. Together, these results show that, in cases where (β, π) is unrestricted, the class of conditional tests has attractive properties. Within this class, however, there remains a question of what test statistics $s(\beta_0)$ to use. In the homoskedastic case, we recommend using the likelihood ratio statistic proposed by Moreira (2003). In the nonhomoskedastic case, however, the literature has not yet converged on a recommendation other than to use one of several procedures that are efficient under strong instruments.

5.2.1. Tests for the homoskedastic case.

A wide variety of test statistics have been proposed in the literature. Kleibergen (2002) proves that a particular score statistic (Breusch & Pagan 1980) has correct size in the model with homoskedastic errors, while in the same model, Moreira (2003) proposes the general conditioning approach for homoskedastic models and notes that both the AR test and Kleibergen's (2002) score test are conditional tests [trivially, since their conditional critical values do not depend on $D(\beta_0)$]. Moreira (2003) further proposes conditional Wald and likelihood ratio tests based on comparing the Wald and likelihood ratio statistics to a conditional critical value. Unlike AR, the score and likelihood ratio statistics depend on both $(\hat{\delta}, \hat{\pi})$, and conditional tests based on these statistics are efficient in the well-identified case.

Andrews et al. (2006) find that the conditional likelihood ratio (CLR) test of Moreira (2003) has very good power properties in the homoskedastic case with a single endogenous variable. The Kronecker product structure of the variance matrix Σ in this setting means that the problem is unchanged by linear transformations of the instruments. It is therefore natural to limit attention to tests that are likewise invariant, in the sense that their value is unchanged by linear transformations of the instruments. Andrews et al. (2006) show, however, that the power of such invariant tests depends only on the correlation between the errors (U, V), the (variance-normalized) length of the first-stage π, and the true parameter value β. Imposing an additional form of invariance to limit attention to two-sided tests, Andrews et al. (2006) show numerically that the CLR test has power close to the upper bound for the power of any invariant similar test over a wide range of parameter values, where the calculation is made feasible by the low dimension of the invariant parameter space. Andrews et al. (2008) extend this result by showing that the power envelope for invariant nonsimilar tests is close to that for invariant similar tests, and thus that (*a*) there is not a substantial power cost to imposing similarity in the homoskedastic setting if one limits attention to invariant tests, and (*b*) that the CLR test performs well even in comparison to nonsimilar tests. Building on these results, Mikusheva (2010) proves a near-optimality property for CLR confidence sets. Andrews et al. (2019) add a note of caution, showing that there exist parameter values not explored by Andrews et al. (2006) where the power of the CLR test is further from the power envelope but still recommend the CLR test for the homoskedastic, single endogenous regressor setting.

5.2.2. Tests for the nonhomoskedastic case.

The simplifications obtained using Kronecker structure of Σ are no longer available in the nonhomoskedastic case, introducing substantial complications.

Motivated by the positive results for the CLR test, several authors have explored analogs and generalizations of the CLR test for nonhomoskedastic settings. In a working paper, Andrews et al. (2004) (for the published version of this paper, see Andrews et al. 2006), introduce a version of the CLR test applicable to the nonhomoskedastic case, while Kleibergen (2005) introduces the conditioning statistic $D(\beta_0)$ for the nonhomoskedastic case and develops score and quasi-CLR statistics applicable in this setting. Andrews & Guggenberger (2015) introduce two alternative quasi-CLR tests for nonhomoskedastic settings that allow a singular covariance matrix Σ. Andrews (2016) studies tests based on linear combinations of AR and score statistics, noting that the CLR test can be expressed in this way. Finally, Moreira & Moreira (2015) and Andrews & Mikusheva (2016) introduce a direct generalization of the CLR test to settings with nonhomoskedastic errors, which again compares the likelihood ratio statistic to a conditional critical value.

All of these extensions of the CLR test are efficient under strong identification, and all but the proposal of Andrews (2016) reduce to the CLR test of Moreira (2003) in the homoskedastic, single endogenous variable setting where the results of Andrews et al. (2006) apply. At the same time, however, while these generalizations are intended for the nonhomoskedastic case, evidence on their performance in the weakly identified case has largely been limited to simulation results.

To derive tests with provable optimality properties in the weakly identified nonhomoskedastic case, a recent literature has focused on optimizing weighted average power, meaning power integrated with respect to weights on (β, π). Specifically the similar test maximizing weighted average power with respect to the weights ν, $\int E_{\beta,\pi}[\phi]\mathrm{d}\nu(\beta, \pi)$, rejects when

$$s(\beta_0) = \int f(\hat{\delta}, \hat{\pi}; \beta, \pi)\mathrm{d}\nu(\beta, \pi) / f(\hat{\delta}, \hat{\pi}|D(\beta_0); \beta_0)$$

exceeds its conditional critical value. Intuitively, this weighted average power optimal test rejects when the observed data are sufficiently more likely to have arisen under the weighted alternative $H_1 : \beta \neq \beta_0$, weighted by ν, than under the null $H_0 : \beta = \beta_0$. As this description suggests, the choice of the weight ν plays an important role in determining the power and other properties of the resulting test, although the use of conditional critical values ensures size control for all choices of ν.

Moreira & Moreira (2013) and Montiel Olea (2018) show that weighted average power optimal similar tests can attain essentially any admissible power function through an appropriate choice of weights. Montiel Olea (2018) further proposes a particular weight ν for the homoskedastic case, while Moreira & Moreira (2015) show that, unless the weights are chosen carefully, weighted average power optimal similar tests may have poor power even in the homoskedastic case, and that the problem can be still worse in the nonhomoskedastic case. To remedy this, they modify the construction of weighted average power optimal tests to enforce a sufficient condition for local unbiasedness and show that these tests perform well in simulation and are asymptotically efficient in the case with strong instruments. Finally, Moreira & Ridder (2017) propose weights ν motivated by invariance considerations. They further show that there exist parameter configurations in the nonhomoskedastic case where tests that depend only on the AR and score statistics, like those of Kleibergen (2005) and Andrews (2016), have poor power.

To summarize, in settings with a single endogenous regressor and homoskedastic errors, the literature to date establishes good properties for the CLR test of Moreira (2003). In settings with nonhomoskedastic errors, by contrast, a large number of procedures have been proposed, but a consensus has not been reached on what procedures to use in practice, beyond the recommendation that researchers use procedures that are efficient when the instruments are strong. Consequently, it is not yet clear what procedure(s) to recommend in this case.

5.3. Inference with Multiple Endogenous Regressors

The tests that we discuss above for models with a single endogenous regressor can all be generalized to tests of hypotheses on the $p \times 1$ vector β in settings with multiple endogenous variables (as in 19 of the 230 specifications in our AER sample). By inverting such tests, we can form simultaneous confidence sets for β. Test inversion with multiple endogenous variables becomes practically difficult for moderate- or high-dimensional β, since the number of grid points at which we need to evaluate our test grows exponentially in the dimension (for a discussion of this issue, see Andrews 2016, **supplemental materials**). In contrast, high-dimensional settings do not appear common in practice, and no specification in our AER data has more than four endogenous regressors. It is in any event rare to report confidence sets for the full vector β in multidimensional settings with strong instruments. Instead, it is far more common to report standard errors or confidence sets for one element of β at a time.

Formally, suppose that we decompose $\beta = (\beta_1, \beta_2)$ and are interested in tests or confidence sets for the subvector β_1 alone. This is known as the subvector inference problem. One possibility for subvector inference is the projection method. In the projection method, we begin with a confidence set CS^{β} for the full parameter vector β and then form a confidence set for β_1 by collecting the implied set of values

$$CS^{\beta_1} = \left\{\beta_1 : \text{there exists } \beta_2 \text{ such that } (\beta_1, \beta_2) \in CS^{\beta}\right\}.$$

This is called the projection method because we can interpret CS^{β_1} as the projection of CS^{β} onto the linear subspace corresponding to β_1. The projection method was advocated for the weak instruments problem by Dufour (1997), Dufour & Jasiak (2001), and Dufour & Taamouti (2005). Dufour & Taamouti (2005) derive analytic expressions for projection-based confidence sets using the AR statistic in the homoskedastic case.

Unfortunately, the projection method frequently suffers from poor power. When used with the AR statistic, for example, we can interpret the projection method as minimizing $AR(\beta_1, \beta_2)$ with respect to the nuisance parameter β_2 and then comparing $\min_{\beta_2} AR(\beta_1, \beta_2)$ to the same χ^2_k critical value that we would have used without minimization. As a result, projection-method confidence sets often cover the true parameter value with probability strictly higher than the nominal level and are thus conservative.

If the instruments are strong for the purposes of estimating β_2 (so that, if β_1 were known, estimation of β_2 would be standard), then these problems have a simple solution: We can reduce our degrees of freedom to account for minimization over the nuisance parameter. Results along these lines for different tests are discussed by Stock & Wright (2000), Kleibergen (2005), and Andrews & Mikusheva (2016).

If we cannot assume that β_2 is strongly identified, then matters are unfortunately more complicated. Guggenberger et al. (2012) show that, in the setting with homoskedastic errors, one can reduce the degrees of freedom for the AR statistic to mitigate projection conservativeness (using a $\chi^2_{k-p_2}$ critical value for p_2 the dimension of β_2), and Guggenberger et al. (2019) propose a further modification to improve power. However, Guggenberger et al. (2012) show that the analog of their result fails for the score statistic of Kleibergen (2002). Moreover, Lee (2015) shows that even the results of Guggenberger et al. (2012) for the AR statistic do not extend to the general nonhomoskedastic case.

To improve the power of the projection method without assuming that the nuisance parameter β_2 is strongly identified, Chaudhuri & Zivot (2011) propose a modified projection approach that chooses the initial confidence set CS^{β} to ensure improved performance for CS^{β_1} in the case with strong instruments. In particular, Chaudhuri & Zivot (2011) base CS^{β} on the combination of a modified score statistic with an AR statistic and show that the resulting CS^{β_1} comes arbitrarily close to efficiency in the case with strong instruments. Andrews (2018) proposes a variant of this approach for constructing confidence sets for functions $f(\beta)$ of the parameter vector other than subvectors, while Andrews (2017) generalizes the work of Chaudhuri & Zivot (2011) in several directions, introducing a variety of test statistics and deriving confidence sets that are asymptotically efficient in the strongly identified case. Finally, Zhu (2015) introduces a Bonferroni approach for subvector inference that provides an alternative to projection.

5.4. Two-Step Confidence Sets

Weak-instrument-robust confidence sets are not widely reported in practice. For instance, only two papers in our AER sample reported robust confidence sets. When such confidence sets are reported, it often appears to be because the authors have uncovered evidence that their instruments are weak. For example, in a survey of 35 empirical papers that reported confidence sets based on Moreira (2003), Andrews (2018) finds that 29 had at least one specification reporting a first-stage F-statistic smaller than 10.

Used in this way, robust confidence sets may act as an alternative to dropping specifications altogether, which, as discussed in Section 4.3, can result in large size distortions. In particular, one can consider constructing a two-step confidence set, where one first assesses instrument strength and then reports conventional confidence sets if the instruments appear strong and a robust confidence set if they appear weak. As discussed by Andrews (2018), the results of Stock & Yogo (2005) imply bounds on the size of two-step confidence sets based on the first-stage F-statistic in homoskedastic or just-identified settings. In overidentified nonhomoskedastic settings, by contrast, Andrews (2018) shows that two-step confidence sets based on the robust first-stage F-statistic F_{R} and conventional cutoffs can have large size distortions. To address this, Andrews (2018) proposes an approach to detecting weak instruments by comparing robust and nonrobust confidence sets. This approach controls coverage distortions for two-step confidence sets in both homoskedastic and nonhomoskedastic settings, including in cases with multiple endogenous variables.

The implications of the negative results of Andrews (2018) for two-step confidence sets based on F_{R} in empirically relevant settings, or for two-step confidence sets based on F_{Eff}, are not clear. To examine this issue, **Figure 6** plots the size of two-step tests based on the effective F-statistic (which use a t-test if $F_{\mathrm{Eff}} > 10$ and an AR test if $F_{\mathrm{Eff}} \leq 10$) against the average effective F-statistic in simulations based on our AER data.

The results of **Figure 6** show that two-step confidence sets based on the effective F-statistic have at most mild size distortions in simulations calibrated to our AER data. Specifically, no specification yields size exceeding 10%, and even when we consider upper bounds, no specification yields size exceeding 11.5%.

6. OPEN QUESTIONS

While considerable progress has been made in both detecting weak instruments and developing identification-robust confidence sets, several important open questions remain. As suggested in Section 5, no consensus has been reached on what inference procedures to use in overidentified models with nonhomoskedastic errors. Likewise, existing optimality results for weak-instrument-robust inference on subsets of parameters only address behavior in the strongly identified case.

Simulation results calibrated to our AER sample raise additional questions. First, we found that conventional t-tests appear to perform reasonably well in specifications where the average effective F-statistic is larger than 10, even in overidentified, nonhomoskedastic cases. Likewise, we found that two-step confidence sets based on the effective F-statistic appear to have well-controlled

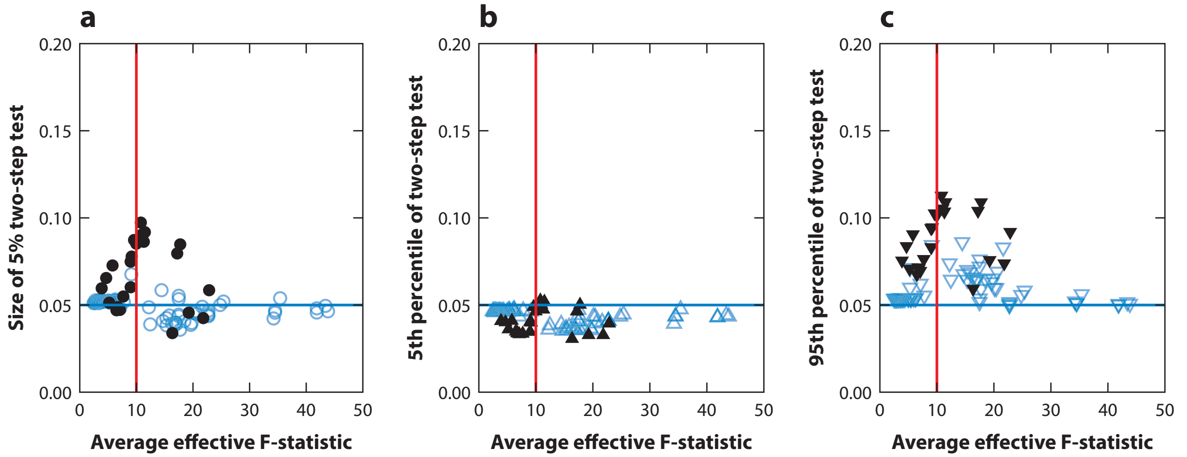

Figure 6

Rejection probability for a nominal 5% two-step test that uses 5% t-test and 5% Anderson-Rubin test when the effective F-statistic is larger than and smaller than 10, respectively, limited to the 106 out of 124 specifications with average effective F smaller than 50. Just-identified specifications are plotted as solid black, while overidentified specifications are plotted as blue outlines. (*a*) Size at parameter values estimated from *American Economic Review* data. (*b*) 5th and (*c*) 95th percentiles of size under the Bayesian exercise described in Section 3. The vertical red line in each panel corresponds to a first-stage F of 10.

size distortions. These results suggest that the effective F-statistic might provide a useful gauge of identification strength in a wider range of cases than is suggested by the current theoretical literature, but a more extensive and formal exploration of whether this is in fact the case, and if so, why, is needed.

Another set of open questions concerns model misspecification. Existing weak-instrument-robust procedures assume that the constant-effect linear IV model holds. If the model is instead misspecified, for example, because the instruments are invalid, then, as noted by Guggenberger (2012), existing weak-instrument-robust confidence sets do not have correct coverage for the true parameter value. Of course, the same is also true for conventional confidence sets with strong but invalid instruments, so this issue is not unique to weak-instrument-robust confidence sets. In overidentified settings with weak instruments, however, the arguments for size control of existing procedures break down even if one considers inference on pseudotrue parameter values (e.g., the population analog of the two-stage least squares or GMM coefficient). This issue is noted and corrected for two-stage least squares estimates in strong-instrument settings with heterogeneous treatment effects by Imbens & Angrist (1994, appendix) and more recently by Lee (2018). To the best of our knowledge, however, analogous results have not been developed for settings with weak instruments.

Concern about model misspecification could also interact with the practice of screening on instrument strength: If one thinks that many instruments used in practice are slightly invalid [in the spirit of, e.g., Conley et al. (2012)], then, while this will result in size distortions, it typically will not qualitatively change results when the instruments are strong. However, when the instruments are weak, even a small degree of instrument invalidity could account for most of the relationship between Z and Y and so lead to qualitatively quite different conclusions. To address this, researchers may wish to limit attention to settings where the instruments are sufficiently strong for them to be confident that results will be qualitatively robust to low levels of instrument invalidity. How to make this argument precise and conduct inference, however, we leave to future work.

Another important open question concerns the validity of the normal approximation to the distribution of the reduced-form and first-stage coefficients. In this review, including in our simulations, we use the model in Equation 6, which takes the reduced-form and first-stage coefficients $(\hat{\delta}, \hat{\pi})$ to be normally distributed with known variance. While this approximation can be justified with asymptotic arguments, whether it is reasonable in a given application is necessarily case specific. Important recent work by Young (2018) casts serious doubt on the quality of this normal approximation in many applications.

Using a sample of studies published in the journals of the American Economic Association that overlaps with but is substantially larger than our AER sample, Young (2018) finds that many reported results are heavily influenced by a small number of observations or clusters. Since the central limit theorem used to derive the limiting normal distribution in Equation 5 for the reduced-form and first-stage coefficients assumes that the influence of each observation is small, this suggests that the normal approximation may be unreasonable. Moreover, Young (2018) notes that variance estimates $\hat{\Sigma}$ for settings with nonhomoskedastic data (which Young calls the non-IID case) can be extremely noisy in finite samples. In simulations that account for these factors, Young finds large size distortions for both conventional and AR tests, with particularly severe distortions for AR tests in overidentified settings. Young (2018) further finds that first-stage F-statistics do not appear to provide a reliable guide to the performance of conventional inference procedures, and that we may spuriously observe large first-stage F-statistics even when the instruments are irrelevant, although he finds somewhat better behavior for the tests of Montiel Olea & Pflueger (2013). To address these issues, Young (2018) suggests using the bootstrap for inference.

We know that bootstrap procedures based on IV estimates or t-statistics are generally invalid when the instruments are weak, and so are not a satisfactory solution in settings with weak instruments. However, appropriately constructed bootstrap procedures based on identification-robust statistics may remain valid. For example, Moreira et al. (2009) show validity of bootstrapped score and AR tests under weak instruments in the homoskedastic case, where it is important for their results that the bootstrap be recentered to ensure that $\hat{\delta} - \hat{\pi}\beta$ has mean zero under the bootstrap distribution. Davidson & MacKinnon (2014) propose additional bootstrap procedures but do not establish their validity when the instruments are weak. We expect that it should be possible to extend the results of Moreira et al. (2009) showing validity of bootstrap-based identification-robust tests to the nonhomoskedastic case and to other identification-robust procedures. At the same time, even when used with identification-robust test statistics, the bootstrap is not a panacea, and Wang & Tchatoka (2018) show that the bootstrap does not ensure size control for subvector inference based on the AR statistic. Given the concerns raised by Young (2018) and the practical importance of the nonhomoskedastic case, such an extension seems like an important topic for future work.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are grateful to Emily Oster and Jesse Shapiro for suggesting the tabulations and simulations based on published results, among other helpful comments, and to Michal Kolesar for bringing the results in the appendix of Imbens & Angrist (1994) to our attention. We are also grateful to the participants in the NBER 2018 Summer Institute Methods Lectures (where we used a draft of this review as lecture notes) and to Donald W.K. Andrews, Adam McCloskey, Marcelo Moreira, and Carolin Pflueger for helpful comments. I.A. gratefully acknowledges support from the National Science Foundation under grant 1654234.

LITERATURE CITED

Anderson T, Rubin H. 1949. Estimators for the parameters of a single equation in a complete set of stochastic equations. *Ann. Math. Stat.* 21:570–82

Andrews D. 2017. *Identification-robust subvector inference*. Discuss. Pap. 2105, Cowles Found., Yale Univ., New Haven, CT

Andrews D, Cheng X, Guggenberger P. 2018. *Generic results for establishing the asymptotic size of confidence sets and tests*. Work. Pap. 1813, Cowles Found. Res. Econ., Yale Univ., New Haven, CT

Andrews D, Guggenberger P. 2009. Asymptotic size and a problem with subsampling and the *m* out of *n* bootstrap. *Econom. Theory* 26:426–68

Andrews D, Guggenberger P. 2015. *Identification- and singularity-robust inference for moment condition models*. Discuss. Pap. 1978, Cowles Found., Yale Univ., New Haven, CT

Andrews D, Guggenberger P. 2017. Asymptotic size of Kleibergen's LM and conditional LR tests for moment condition models. *Econom. Theory* 33:1046–80

Andrews D, Marmer V, Yu Z. 2019. On optimal inference in the linear IV model. *Quant. Econ.* 10(2):457–85

Andrews D, Moreira M, Stock J. 2004. *Optimal invariant similar tests of instrumental variables regression*. Discuss. Pap. 1476, Cowles Found., Yale Univ., New Have, CT

Andrews D, Moreira M, Stock J. 2006. Optimal two-sided invariant similar tests of instrumental variables regression. *Econometrica* 74:715–52

Andrews D, Moreira M, Stock J. 2008. Efficient two-sided nonsimilar invariant tests in IV regression with weak instruments. *J. Econom.* 146:241–54

Andrews I. 2016. Conditional linear combination tests for weakly identified models. *Econometrica* 84:2155–82

Andrews I. 2018. Valid two-step identification-robust confidence sets for GMM. *Rev. Econ. Stat.* 100:337–48

Andrews I, Armstrong TB. 2017. Unbiased instrumental variables estimation under known first-stage sign. *Quant. Econ.* 8:479–503

Andrews I, Mikusheva A. 2016. Conditional inference with a functional nuisance parameter. *Econometrica* 84:1571–612

Angrist J, Krueger A. 1991. Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.* 106:979–1014

Baum C, Schaffer M, Stillman S. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata J.* 7:465–506

Bound J, Jaeger D, Baker R. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *J. Am. Stat. Assoc.* 90:443–50

Breusch T, Pagan A. 1980. The Lagrange multiplier test and its applications to model specifications in econometrics. *Econometrica* 47:239–53

Chaudhuri S, Zivot E. 2011. A new method of projection-based inference in GMM with weakly identified nuisance parameters. *J. Econom.* 164:239–51

Chernozhukov V, Jansson M, Hansen C. 2009. Admissible invariant similar tests for instrumental variables regression. *Econom. Theory* 25:806–18

Conley T, Hansen C, Rossi P. 2012. Plausibly exogenous. *Rev. Econ. Stat.* 94:260–72

Cragg J, Donald S. 1993. Testing identifiability and specification in instrumental variable models. *Econom. Theory* 9:222–40

Davidson R, MacKinnon J. 2014. Bootstrap confidence sets with weak instruments. *Econom. Rev.* 33:651–75

Dufour J. 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65:1365–87

Dufour J, Jasiak J. 2001. Finite sample limited information inference methods for structural equations and models with generated regressors. *Int. Econ. Rev.* 42:815–44

Dufour J, Taamouti M. 2005. Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica* 73:1351–65

Favara G, Imbs J. 2015. Credit supply and the price of housing. *Am. Econ. Rev.* 105:958–92

Fieller E. 1954. Some problems in interval estimation. *J. R. Stat. Soc. B* 16:175–85

Gleser L, Hwang J. 1987. The nonexistence of 100(1−α)% confidence sets of finite expected diameter in errors-in-variables and related models. *J. Am. Stat. Assoc.* 15:1341–62

Guggenberger P. 2012. On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. *Econom. Theory* 28:387–421

Guggenberger P, Kleibergen F, Mavroeidis S. 2019. A more powerful subvector Anderson Rubin test in linear instrumental variable regression. *Quant. Econ.* 10(2):487–526

Guggenberger P, Kleibergen F, Mavroeidis S, Chen L. 2012. On the asymptotic sizes of subset Anderson–Rubin and Lagrange multiplier tests in linear instrumental variables regression. *Econometrica* 80:2649–66

Hausman JA. 1983. Specification and estimation of simultaneous equation models. In *Handbook of Econometrics*, ed. Z Grilliches, M Intriligator, pp. 391–448. Amsterdam: North-Holland

Hirano K, Porter J. 2015. Location properties of point estimators in linear instrumental variables and related models. *Econom. Rev.* 34:720–33

Hornung E. 2014. Immigration and the diffusion of technology: the Huguenot diaspora in Prussia. *Am. Econ. Rev.* 104:84–122

Imbens G, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–75

Kleibergen F. 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* 70:1781–803

Kleibergen F. 2005. Testing parameters in GMM without assuming they are identified. *Econometrica* 73:1103–23

Kleibergen F. 2007. Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices, and identification statistics. *J. Econom.* 139:181–216

Kleibergen F, Paap R. 2007. Generalized reduced rank tests using the singular value decomposition. *J. Econom.* 133:97–126

Lee J. 2015. *Asymptotic sizes of subset Anderson-Rubin tests with weakly identified nuisance parameters and general covariance structure*. Unpublished manuscript, Mass. Inst. Technol., Cambridge, MA

Lee S. 2018. A consistent variance estimator for 2SLS when instruments identify different LATEs. *J. Bus. Econ. Stat.* 36:400–10

Lehmann E, Romano J. 2005. *Testing Statistical Hypotheses*. Berlin: Springer. 3rd ed.

Magnusson L. 2010. Inference in limited dependent variable models robust to weak identification. *Econom. J.* 13:S56–79

Mariano R, Sawa T. 1972. The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included endogenous variables. *J. Am. Stat. Assoc.* 67:159–63

Mikusheva A. 2010. Robust confidence sets in the presence of weak instruments. *J. Econom.* 157:236–47

Montiel Olea J. 2018. *Admissible, similar tests: a characterization*. Unpublished manuscript, Columbia Univ., New York

Montiel Olea J, Pflueger C. 2013. A robust test for weak instruments. *J. Bus. Econ. Stat.* 31:358–69

Moreira H, Moreira M. 2013. *Contributions to the theory of optimal tests*. Unpublished manuscript, FGV/EPGE, Rio de Janeiro

Moreira H, Moreira M. 2015. *Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors*. Work. Pap. CWP25/16, Cent. Microdata Methods Pract., London

Moreira M. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71:1027–48

Moreira M. 2009. Tests with correct size when instruments can be arbitrarily weak. *J. Econom.* 152:131–40

Moreira M, Porter J, Suarez G. 2009. Bootstrap validity for the score test when instruments may be weak. *J. Econom.* 149:52–64

Moreira M, Ridder G. 2017. *Optimal invariant tests in an instrumental variables regression with heteroskedastic and autocorrelated errors*. Unpublished manuscript, FGV/EPGE, Rio de Janeiro

Nagar A. 1959. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 27:575–95

Nelson C, Startz R. 1990a. The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one. *J. Bus.* 63:5125–40

Nelson C, Startz R. 1990b. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58:967–76

Sanderson E, Windmeijer F. 2016. A weak instrument f-test in linear IV models with multiple endogenous variables. *J. Econom.* 190:212–21

Sawa T. 1969. The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *J. Am. Stat. Assoc.* 64:923–37

Staiger D, Stock J. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65:557–86

Stock J, Wright J. 2000. GMM with weak identification. *Econometrica* 68:1055–96

Stock J, Yogo M. 2005. Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. DWK Andrews, JH Stock, pp. 80–108. Cambridge, UK: Cambridge Univ. Press

Wang W, Tchatoka FD. 2018. On bootstrap inconsistency and Bonferroni-based size-correction for the subset Anderson–Rubin test under conditional homoskedasticity. *J. Econom.* 207:188–211

Young A. 2014. Structural transformation, the mismeasurement of productivity growth, and the cost disease of services. *Am. Econ. Rev.* 104:3635–67

Young A. 2018. *Consistency without inference: instrumental variables in practical application*. Unpublished manuscript, London School Econ.

Zhu Y. 2015. *A new method for uniform subset inference of linear instrumental variables models*. Unpublished manuscript, Univ. Oregon, Eugene

Zivot E, Startz R, Nelson CR. 1998. Valid confidence regions and inference in the presence of weak instruments. *Int. Econ. Rev.* 39:1119–46