Federal Reserve Bank of Dallas

Online Appendix to A Robust Test for Weak Instruments for 2SLS with Multiple Endogenous Regressors

Daniel J. Lewis and Karel Mertens

Working Paper 2208 Appendix **September 2024**

Research Department

https://doi.org/10.24149/wp2208app

Working papers from the Federal Reserve Bank of Dallas are preliminary drafts circulated for professional comment. The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Dallas or the Federal Reserve System. Any errors or omissions are the responsibility of the authors.

A Robust Test for Weak Instruments With Multiple Endogenous Regressors

Daniel Lewis Karel Mertens

ONLINE APPENDIX

E Additional Proofs

E.1 Proof of Proposition 3.

The cdf under the Imhof (1961) approximation in Definition 5 is

(E.34)

$$\mathcal{F}_I(x;\kappa_1,\kappa_2,\kappa_3) = \mathcal{F}_{\chi^2_\nu}((x-\kappa_1)4\omega+\nu) = \int_{\kappa_1-\nu(4\omega)^{-1}}^{x} f(z)dz \quad \text{, where}$$

$$\nu = 8\kappa_2\omega^2 \quad ; \quad \omega = \kappa_2/\kappa_3 \quad ; \quad f(z) = \left(1+\frac{z-\kappa_1}{2\kappa_2\omega}\right)^{\nu/2-1} e^{-\frac{\nu}{2}\left(1+\frac{z-\kappa_1}{2\kappa_2\omega}\right)} \frac{(\nu/2)^{\nu/2-1}\omega}{2^{\nu/2-2}\Gamma(\nu/2)}.$$

The pdf $f(z)$ has a mode at $z^m = \kappa_1 - (2\omega)^{-1}$ if $\nu \geq 2$, and at zero otherwise. The critical value associated with the upper α-percentile, $x(\alpha)$, is implicitly defined by $\alpha = \int_{x(\alpha)}^{\infty} f(z)dz$. To find the largest possible critical value among all possible distributions, we solve the following optimization problem:

$$\max_{\kappa_1,\kappa_2,\kappa_3} \; x(\alpha) \;\; \text{s.t.} \;\; \kappa_n \leq \bar{\kappa}_n \text{ for } n = 1,2,3\ . \tag{E.35}$$

Consider the Kuhn-Tucker conditions

$$\int_{x(\alpha)}^{\infty} \frac{\partial f(z)}{\partial \kappa_n} dz = \mu_n, \tag{E.36}$$

together with $\mu_n \geq 0$, $n = 1,2,3$, the constraints and the complementary slackness conditions, where μ_n are the multipliers times $f(x(\alpha)) > 0$. The Kuhn-Tucker conditions follow from the implicit function theorem and Leibniz's rule: $1 = -f(x(\alpha))\frac{\partial x(\alpha)}{\partial y} + \int_{x(\alpha)}^{\infty} \frac{\partial f(z)}{\partial y} dz \Rightarrow \frac{\partial x(\alpha)}{\partial y} = \int_{x(\alpha)}^{\infty} \frac{\partial f(z)}{\partial y} dz / f(x(\alpha))$ with $f(x(\alpha)) > 0$ for $\alpha \in (0,1)$.

The partial derivatives are

$$\frac{\partial f(z)}{\partial \kappa_1} = \frac{1 + (z - \kappa_1)2\omega}{2\kappa_2\omega} \left(1 + \frac{z - \kappa_1}{2\kappa_2\omega}\right)^{-1} f(z), \tag{E.37}$$

$$\frac{\partial f(z)}{\partial \kappa_2} = \frac{f(z)}{\kappa_2} G_1\left((z - \kappa_1)4\omega + \nu\right), \tag{E.38}$$

$$\frac{\partial f(z)}{\partial \kappa_3} = \frac{f(z)}{\kappa_3} G_2\left((z - \kappa_1)4\omega + \nu\right), \tag{E.39}$$

where

$$G_1(y) = -\frac{1}{2}\left(y - 2\nu(\nu - 2)/y + \nu\right) + 3/2(\ln(y/2) - \psi(\nu/2))\nu, \tag{E.40}$$

$$G_2(y) = \frac{1}{2}\left(y - \nu(\nu - 2)/y\right) - (\ln(y/2) - \psi(\nu/2))\nu, \tag{E.41}$$

and $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function (the logarithmic derivative of the gamma function $\Gamma(x)$). From Alzer (1997) (Equation 2.2), we know that

$$1/\nu < \ln(\nu/2) - \psi(\nu/2) < 2/\nu. \tag{E.42}$$

For $n = 1$, the LHS of (E.36) is always positive to the right of the mode, which means the constraint on the mean ($n = 1$) is always binding. The Alzer bounds imply that in the right tail of any optimal distribution, the LHS of (E.36) is always strictly positive for $n = 2, 3$, which means that the constraints are also binding as long as α is sufficiently small.

E.2 Proof of Corollary 1

The absolute bias criterion for the j-th element is

$$B^j_{abs} = \sqrt{e_j^{N\prime}\Sigma_v e_j^N}\, e_j^{N\prime} E\left[\beta^*_{2SLS}\right]/\sigma_u \tag{E.43}$$

Using the definitions of $\mathbf{h}$ and ρ_{abs} in the main text,

$$B^j_{abs} = \sqrt{e_j^{N\prime}\Sigma_v e_j^N}\, e_j^{N\prime}\Phi^{-\frac{1}{2}}\mathbf{h}\rho_{abs} \tag{E.44}$$

Defining the $N \times 1$ vector $\mu_j = \sqrt{e_j^{N\prime}\Sigma_v e_j^N}\,\Phi^{-\frac{1}{2}} e_j^N$ and using the Nagar approximation of $\mathbf{h}$ yields the Nagar bias

$$B^j_{abs,n} = \mu_j'\mathbf{h}_n\rho_{abs} \tag{E.45}$$

Using the definition of ρ_{abs} in (B.9), and following the same steps as in (B.10),

$$\sup_{\beta\in\mathbb{R}^N} B^j_{abs,n} = K^{-\frac{1}{2}}||\mu_j'\mathbf{h}_n\Psi_{abs}||_2 \tag{E.46}$$

Using the same arguments as in the proof of Theorem 1,

$$\sup_{\mathcal{D}_\Lambda\geq\lambda_{\min}I_N} K^{-\frac{1}{2}}||\mu_j'\mathbf{h}_n\Psi_{abs}||_2 = K^{-\frac{1}{2}}\lambda_{\min}^{-1}||\mu_j'Q_\Lambda M_1(Q_\Lambda\otimes L_0\otimes L_0)M_2\Psi_{abs}||_2 \tag{E.47}$$

where all objects are defined as in Theorem 1 and Appendix B.
Since $\sup_{L_0\in\mathbb{O}^{N\times K}}||M_1(Q_\Lambda\otimes L_0\otimes L_0)M_2\Psi_{abs}||_2 = \sup_{L_0\in\mathbb{O}^{N\times K}}||M_1(I_N\otimes L_0\otimes L_0)M_2\Psi_{abs}||_2$ for any $Q_\Lambda\in\mathbb{O}^{N\times N}$ and since it is possible to choose Q_Λ such that the orthonormal vector $Q_\Lambda'\mu_j/||\mu_j||_2$ selects the largest singular value of $M_1(I_N\otimes L_0\otimes L_0)M_2\Psi_{abs}$, it follows that a sharp upper bound on the Nagar bias is

$$\begin{aligned} B^{j*}_{abs,n}(\mathbf{W},\lambda_{\min}) &= ||\mu_j||_2\times\lambda_{\min}^{-1}K^{-\frac{1}{2}}\sup_{L_0\in\mathbb{O}^{N\times K}}\{||M_1(I_N\otimes L_0\otimes L_0)M_2\Psi_{abs}||_2\} \\ &= ||\mu_j||_2\times B^*_{abs,n} \end{aligned} \tag{E.48}$$

where $B^*_{abs,n}$ is defined in Appendix B and $||\mu_j||_2 = \sqrt{e_j^{N\prime}\Sigma_v e_j^N}\ ||\Phi^{-\frac{1}{2}}e_j^N||_2$ is the constant in the adjustment to the tolerance level for the absolute bias criterion in Corollary 1.

For the relative bias criterion, the worst-case 2SLS bias benchmark changes due to the change in weights. In particular, using (9) and evaluating the bias criterion gives the worst-case benchmark

$$\sqrt{\mathrm{Tr}(\mathbf{S}_1)}\left(\sup_{||x||\leq 1} x'\Phi^{-\frac{1}{2}}e_j^N e_j^{N\prime}\Phi e_j^N e_j^{N\prime}\Phi^{-\frac{1}{2}}x\right)^{\frac{1}{2}} = \sqrt{\mathrm{Tr}(\mathbf{S}_1)}\sqrt{e_j^{N\prime}\Phi e_j^N}||\Phi^{-\frac{1}{2}}e_j^N||_2 \tag{E.49}$$

Therefore

$$\begin{aligned} B^{j*}_{rel,n}(\mathbf{W},\lambda_{\min}) &= \frac{||\mu_j||_2}{\sqrt{e_j^{N\prime}\Phi e_j^N}||\Phi^{-\frac{1}{2}}e_j^N||_2}\times B^{j*}_{rel,n}\ , \\ &= B^{j*}_{rel,n}\ . \end{aligned} \tag{E.50}$$

where $B^*_{rel,n}$ is defined in Appendix B, such that no adjustment is required to the tolerance level for the relative bias criterion.

E.3 Proof of Corollary 2

First, note that the Schur complement of $Y_j'P_ZY_j$ in $Y'P_ZY$ is given by

$$Y_j'P_ZY_j - Y_j'P_ZY_{-j}(Y_{-j}'P_ZY_{-j})^{-1}Y_{-j}'P_ZY_j = Y_j^{\perp\prime}P_{Z^\perp}Y_j^\perp. \tag{E.51}$$

Using standard formulas for the inverse of a partitioned matrix and properties of projection matrices, partition $\hat{\beta}_{2SLS} - \beta = (Y'P_ZY)^{-1}YP_Zu$ as

$$\begin{aligned} \hat{\beta}_{2SLS,j} - \beta_j &= (Y_j^{\perp\prime}P_{Z^\perp}Y_j^\perp)^{-1}(Y_j - Y_{-j}\hat{\delta})'P_Zu \ , \\ \hat{\beta}_{2SLS,-j} - \beta_{-j} &= (Y_{-j}'P_ZY_{-j})^{-1}Y_{-j}'P_Zu - \hat{\delta}\left(\hat{\beta}_{2SLS,j} - \beta_j\right) \ , \end{aligned} \tag{E.52}$$

where $u^\perp = M_{\hat{Y}_{-j}}u$ and $\hat{\delta} = (Y_{-j}'P_ZY_{-j})^{-1}Y_{-j}'P_ZY_j$. Since $\hat{\delta} \xrightarrow{p} \delta$ and $(Y_{-j}'P_ZY_{-j})^{-1}Y_{-j}'P_Zu \xrightarrow{p} 0$,

$$\hat{\beta}_{2SLS} - \beta \xrightarrow{d} \beta^*_{2SLS} = \tilde{\delta}\beta^*_{2SLS,j} \ , \text{where } \tilde{\delta}_j = 1 \ , \tilde{\delta}_{-j} = -\delta. \tag{E.53}$$

Under Assumption 3, the bias criterion is therefore

$$B_i = |E[\beta^*_{2SLS,j}]| \times ||\Xi_i^{\frac{1}{2}}\tilde{\delta}||_2/\sqrt{b_i} \ . \tag{E.54}$$

The bias in the transformed single-regressor regression for j is

$$B_i^\perp = |E[\beta^*_{2SLS,j}]| \times \sqrt{\Xi_i^\perp}/\sqrt{b_i^\perp} \ . \tag{E.55}$$

where $\Xi_i^\perp$ and $b_i^\perp$ are the weighting and scaling for the transformed regression with a single regressor.

Therefore,

$$B_i = B_i^\perp \times \frac{||\Xi_i^{\frac{1}{2}}\tilde{\delta}||_2}{\sqrt{\Xi_i^\perp}} \frac{\sqrt{b_i^\perp}}{\sqrt{b_i}} \ . \tag{E.56}$$

Since $b_{abs} = b_{abs}^\perp = \sigma_u^2$ and $\Xi_{abs} = \Sigma_v$ and $\Xi_{abs}^\perp = \tilde{\delta}'\Sigma_v\tilde{\delta}$,

$$B_{abs} = B_{abs}^\perp \times \frac{||\Sigma_v^{\frac{1}{2}}\tilde{\delta}||_2}{\sqrt{\tilde{\delta}'\Sigma_v\tilde{\delta}}} = B_{abs}^\perp \ . \tag{E.57}$$

As the bias B_{abs} in the original regression with multiple endogenous regressors is identical to the bias $B_{abs}^\perp$ in the transformed regression with a single regressor , the full-vector weak instruments test can be based on a weak instruments test in the transformed regression as stated in part (i).

The absolute bias criterion for the j-th element in (E.58) is a simple

reweighting of $B^{\perp}_{abs}$,

$$B^{j}_{abs} = B^{\perp}_{abs} \times \frac{\sqrt{e_j^{N\prime}\Sigma_v e_j^N}}{\sqrt{\tilde{\delta}'\Sigma_v\tilde{\delta}}} \tag{E.58}$$

such that part (ii) and the stated adjustment to the tolerance level follow immediately.

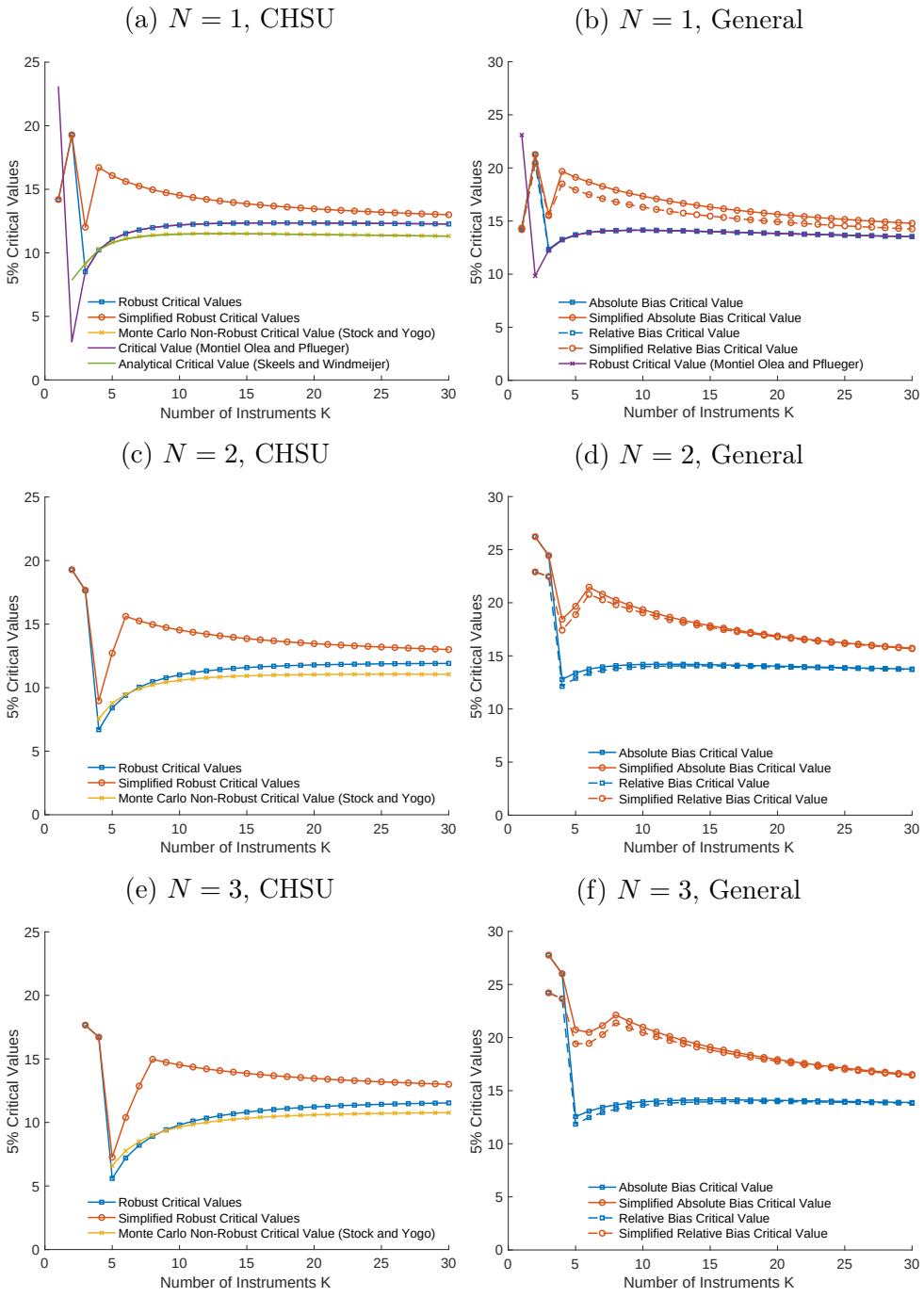
F Illustrative Comparison of Critical Values Across Tests

This Appendix discusses the relationship between the critical values of our generalized robust weak instruments tests and others in the literature. Figure F.1 shows critical values for $\alpha = 0.05$ and $\tau = 0.10$ across several tests and for a range of specifications with $N = 1$, 2, or 3 endogenous regressors.

We first discuss the panels for CHSU models in the first column of Figure F.1. With conditionally homoskedastic and serially uncorrelated errors, the two measures of the bias – absolute bias and relative bias – are identical, and the critical values only depend on N and K. The figures in the left column report the critical values for our robust test, as well as the alternative conservative simplified values. Each panel also plots the critical values from the Stock and Yogo (2005) tables for comparison, which are available for $K > N+1$. The top left panel for $N = 1$ also reports the critical values from the Montiel Olea and Pflueger (2013) test for $K \geq 1$, as well as the analytical critical values derived for $K > 1$ by Skeels and Windmeijer (2018).

For CHSU models with $K > N + 1$, the differences in the critical values are relatively small. These differences arise exclusively because of the use of a Nagar approximation (our test and Montiel Olea and Pflueger 2013), Monte Carlo integration (Stock and Yogo 2005), or analytical expressions (Skeels and Windmeijer 2018). As discussed in detail in Section 2.5, for $N = 1$ and $K = 1$ or $K = 2$, the robust critical values from our test differ from those of Montiel Olea and Pflueger (2013) in the top left panel. When $K = 1$, our critical value is lower because it is based on the median bias, and when $K = 2$, it is larger because we use a more conservative bound. The accuracy problems of not adopting this bound are evident in the figure, as the Montiel Olea and Pflueger (2013) critical value is 3.00 for $K = 2$, whereas the analytical critical value of Skeels and Windmeijer (2018) is 7.85. In the CHSU model, our approach leads to a much more conservative critical value close to 20. Finally, as expected, the simplified critical values are conservative for $K > N+1$. The difference between the simplified critical values and those

FIGURE F.1: Comparison of Critical Values



Notes: The left column reports critical values for $\alpha = 0.05$ and $\tau = 0.10$ for models with conditional homoskedasticity and no serial correlation (CSHU) and for various numbers of endogenous regressors (N) and instruments (K). The right column repeats the exercise for models with arbitrary heteroskedasticity and/or autocorrelation. Critical values depend on $\mathbf{W}$ and therefore vary for each application. The figures show averages of over 500 draws of $\mathbf{W}$ as described in Section 4 for illustrative purposes only. For comparison, we plot applicable critical values from Montiel Olea and Pflueger (2013), Stock and Yogo (2005), and Skeels and Windmeijer (2018).

based on the worst-case Nagar bias diminishes as K increases.

The right column in Figure F.1 shows critical values non-CHSU models with general $\mathbf{W}$. As the critical values depend additionally on the covariance matrix $\mathbf{W}$ (and Σ_{wv} for absolute bias), they will be different for each application. To nevertheless give a sense of the critical values that arise in practice, the figures show the average robust critical values across 500 different general covariance matrices drawn randomly from the process described in Section G. The top panel for the model with $N = 1$ additionally reports the robust Montiel Olea and Pflueger (2013) critical values. The averages across DGPs mask considerable variation in critical values across $\mathbf{W}$'s, but they are still useful to illustrate a number of noteworthy features.

As the figures show, the absolute and relative bias criteria now lead to different critical values. On average, the absolute and relative critical values are very similar. The absolute bias critical values are marginally larger on average, and the difference with the relative bias critical values vanishes as K grows large. For $N = 1$ and $K > 2$, our critical values for the relative bias coincide almost exactly with those of Montiel Olea and Pflueger (2013), and for $N = 1$ and $K \leq 2$, they differ for the same reasons discussed above and in Section 2.5. As in the left panel, the simplified critical values are conservative for $K > N + 1$, and the difference with those based on the worst-case Nagar bias vanishes for large K.

G Asymptotic Simulation Study

This section reports the results from simulations in which the random vectors entering $\hat{\beta}_{2SLS}$ are drawn directly from their asymptotic distribution, see Equation (7). These asymptotic simulations complement the finite sample simulations in Section 4 as they make it computationally feasible to verify the performance of the robust tests across a much larger set of models.

We consider models with six different combinations for the number of endogenous variables and the number of instruments: $N = 2$, with $K = 2, 3, 4, 6$, and $N = 3$, with $K = 5, 9$. For each combination of N and K, we consider five million randomly drawn DGPs $\{\beta, C, \mathbf{W}\}$. We first generate 10,000 $\mathbf{W}$ matrices. To do so, we draw a random Σ_{wv} from the standard Wishart distribution with $N + 1$ degrees of freedom and then generate $\mathbf{W}$, conditional on Σ_{wv}, following a similar process to that described for the baseline simulations in the main text. In particular, we assume $\begin{bmatrix} w_t, & v_t' \end{bmatrix}' \sim \mathcal{N}(0, \Sigma_{wv})$ and

$\bar{Z}_t \sim \mathcal{N}(0, Q_t)$, where

$$Q_t = I_K + \Gamma \begin{bmatrix} w_t \\ v_t \end{bmatrix} \begin{bmatrix} w_t & v_t' \end{bmatrix}' \Gamma', \tag{G.59}$$

and

$$Z_t = (\bar{Z}'\bar{Z}/T)^{-\frac{1}{2}} \bar{Z}_t. \tag{G.60}$$

We then evaluate

$$\mathbf{W} = \text{cov}\left(\begin{bmatrix} Z_t w_t \\ \text{vec}(Z_t v_t') \end{bmatrix}, \begin{bmatrix} Z_t w_t \\ \text{vec}(Z_t v_t') \end{bmatrix}' \right) \tag{G.61}$$

analytically using the higher moments of the multivariate Gaussian distribution. This procedure ensures that Σ_{wv} and $\mathbf{W}$ are mutually consistent, as both are needed for our absolute bias test. For each $\mathbf{W}$, we draw 10 different pairs of values for β and directions C_0, defined such that $C = \sqrt{\lambda_{\min}} C_0$. For each of the resulting 100,000 draws, we consider a grid of 50 minimum eigenvalues of the concentration matrix, $\lambda_{\min}$, over a range from 0.25 to 100. To have good coverage of the region where the Nagar bias is maximized for a given $\lambda_{\min}$, half of the draws for β and C_0 are in a neighborhood of the 'worst-case' values $\{\beta^{wc}, C_0^{wc}\}$ where the Nagar bias is at the upper bound for a given $\mathbf{W}$. The other half of the draws for β and C_0 are from a wider region of the parameter space.[1] For each of the resulting five million DGPs, we generate 1000 samples for the random vectors η_1 and η_2 in (7) to draw from the limiting distribution of $g_{\min}$ and obtain the empirical rejection rates of the first-stage tests.

G.1 Accuracy of the Nagar Approximation

Because the Nagar bias is only an approximation of the bias, it is not immediate that the worst-case Nagar bias is an effective upper bound on the bias in

[1] For β, the draws close to the worst-case Nagar bias are $\beta = \beta^{wc} + 0.1\upsilon$ where $\upsilon \sim \mathcal{N}(0, \mathcal{I}_N)$, and the other draws comprise N independent draws from the uniform distribution on $[-100, 100]$. For C_0, we use the reparametrization $\text{vec}(C_0') = \mathbf{S}_2^{\frac{1}{2}} \mathcal{S}^{-1} \sqrt{K}\, \text{vec}(L_0' \mathcal{D}_{\Lambda_0}^{\frac{1}{2}} Q_\Lambda')$, where L_0 is an orthonormal matrix as in Theorem 1, and $\lambda_{\min} D_{\Lambda_0}$ and Q_Λ contain the eigenvalues and -vectors of the concentration matrix Λ. The orthonormal matrix Q_Λ is always drawn from the Haar distribution. For the draws close to the worst-case Nagar bias, we set $L_0 = ((L_0^{wc} + \xi)(L_0^{wc} + \xi)')^{-1} (L_0^{wc} + \xi)$ where the elements of ξ are drawn independently from a uniform distribution on $[-0.1, 0.1]$ and L_0^{wc} is the orthonormal matrix that maximizes $\mathcal{B}_\rangle(\mathbf{W})$ in Theorem 1. The other draws of L_0 are from the Haar distribution. Finally, the nonzero diagonal elements of $\mathcal{D}_{\Lambda_0}$ are generated as $1 + 0.1\upsilon$ with $\upsilon \sim \chi^2(1)$ for the draws close to the worst case, and from a $\chi^2(1)$ distribution in the other draws. In both cases, the draws are normalized such that the smallest diagonal element of $\mathcal{D}_{\Lambda_0}$ is unity.

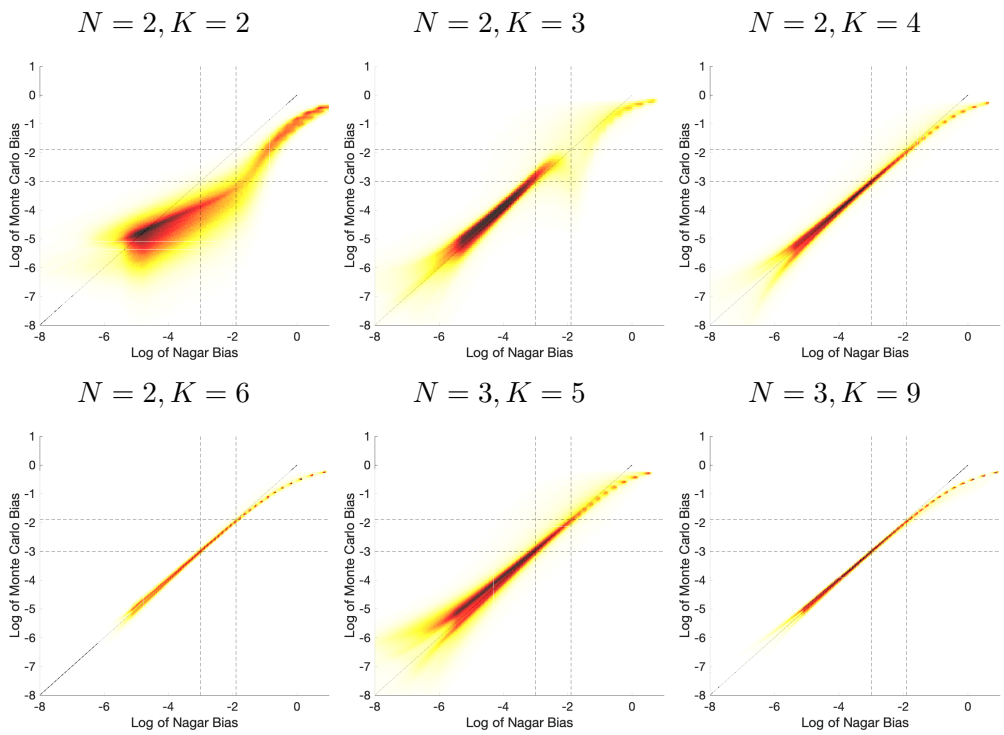
practice. We therefore first assess numerically whether the worst-case Nagar bias (which depends on $\lambda_{\min}$ and $\mathbf{W}$) is a valid bound for the bias (which depends on β, C, and $\mathbf{W}$) across the DGPs. To do so, we evaluate the bias using Monte Carlo integration for each DGP. More specifically, to compute the Monte Carlo bias we evaluate the bias criterion in Definition 2 by replacing $E[\beta^*_{2SLS}]$ with simulated sample averages of β^*_{2SLS}. As discussed in Section 2.5, in just-identified models, the test is based on the median bias rather than the mean, and so in those models, the Monte Carlo bias is the median across simulated samples. We restrict attention to the results for the absolute bias criterion for brevity, but those for the relative bias are very similar.

For the models with $K > N+1$, we find in the simulations that the worst-case Nagar bias is a highly effective upper bound on the Monte Carlo bias at a conventional bias tolerance level of $\tau = 0.10$. Across all four specifications with $K > N+1$, the Monte Carlo bias exceeds 0.10 in fewer than 0.0001% of the DGPs for which the worst-case Nagar bias is smaller than 0.10. In the two models with $K \leq N+1$, on the other hand, the Monte Carlo bias exceeds 0.10 in 0.0004% ($K = N = 2$) and 0.33% ($N = 2, K = 3$) of the DGPs with worst-case Nagar bias less than 0.10. These relatively more frequent failures of the Nagar approximation in models with $K \leq N+1$ is why we adopt the more conservative upper bound in our testing procedure whenever $K \leq N+1$.[2] Using these more conservative bounds, the bound never falls below $\tau = 0.10$ when the Monte Carlo bias is above $\tau = 0.10$ for either $K \leq N+1$ DGP.

For a broader perspective on the quality of the Nagar approximation, Figure G.2 plots the log of the Monte Carlo bias against the log of the Nagar bias as in Definition 4 (i.e., not the worst-case Nagar bias but the Nagar bias evaluated at the true parameters). As the figure shows, the relationship between the Nagar and Monte Carlo bias becomes considerably stronger with the degree of overidentification. The dashed lines in Figure G.2 mark a bias range of 0.05 to 0.15. This is likely to be the most relevant range in practice, and the accuracy of the Nagar bias in this range is, therefore, the most important. The figure shows that over that range, the relationship between the Monte Carlo bias and the Nagar bias is very strong in all models with $K > N+1$. For models with $K \leq N+1$, Figure G.2 illustrates the limitations of the Nagar approximation, as the relationship with the Monte Carlo bias is meaningfully weaker.

[2]Section G.3 shows that the failures of the Nagar approximation for $K = N+1$ become even more dramatic in homoskedastic models, as also discussed in Section 2.5.

FIGURE G.2: Comparison of Nagar Bias and Monte Carlo Bias, Absolute Bias



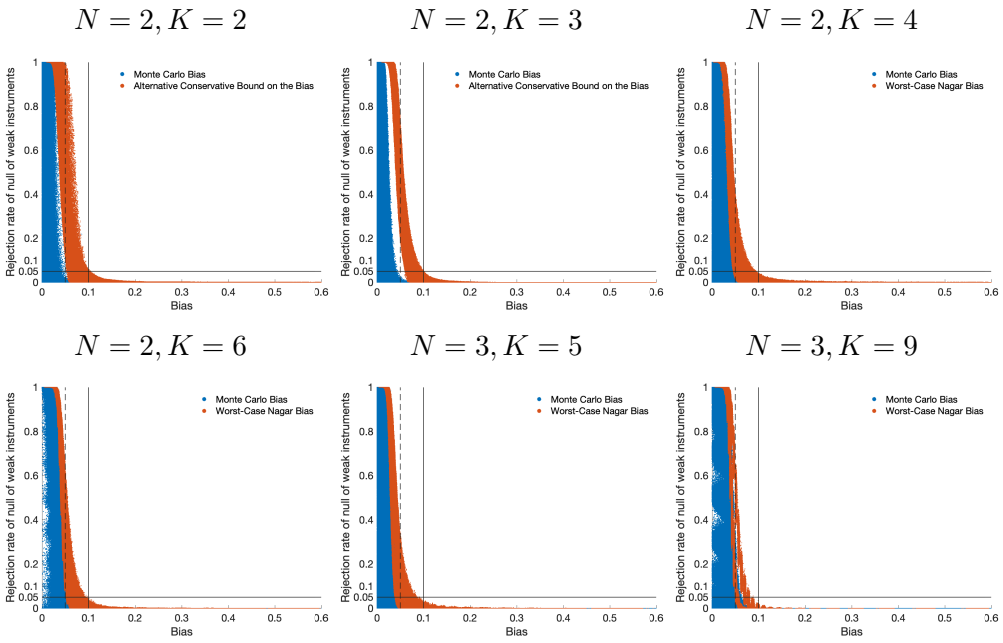
Notes: For each specification, we consider five million DGPs as described in the main text. For each DGP, we take 1000 samples and compute the Monte Carlo bias by numerical integration. The figure plots the (log of) Monte Carlo bias against the Nagar bias, $B_{abs,n}$. For $N = 2, K = 2$, both consider the median bias. The heatmap indicates the density of DGPs with a particular combination of Nagar and Monte Carlo biases. The dashed horizontal and vertical lines indicate bias levels of 0.05 and 0.15 to demarcate the typically relevant region for first-stage tests.

G.2 Size and Power of the Robust First-Stage Test

Figures G.3 and G.4 present scatter plots of the empirical rejection rates of the absolute and relative bias tests of Section 2 as a function of bias across the five million DGPs. In all cases, $\tau = 0.10$ (vertical lines) and $\alpha = 0.05$ (horizontal lines). The red dots plot the rejection rates as a function of the worst-case Nagar bias (when $K > N + 1$) or the more conservative bound (when $K \leq N + 1$), evaluated at the values of $\lambda_{\min}$ and $\mathbf{W}$ in each DGP. For illustration, the blue dots also plot the rejection rates against the Monte Carlo bias given the values of $\{\beta, \sqrt{\lambda_{\min}} C_0, \mathbf{W}\}$ in each DGP (but not taking the worst case over all possible β and C_0). Of course, the test will have weakly lower rejection rates when plotted against these values since the null hypothesis pertains to upper bounds on (the Nagar approximation of) the bias.

If the test is perfectly sized, the empirical rejection rates should equal the nominal size of $\alpha = 0.05$ when the worst-case Nagar bias is precisely $\tau = 0.10$.

FIGURE G.3: Size and Power of the First-Stage Test, Absolute Bias

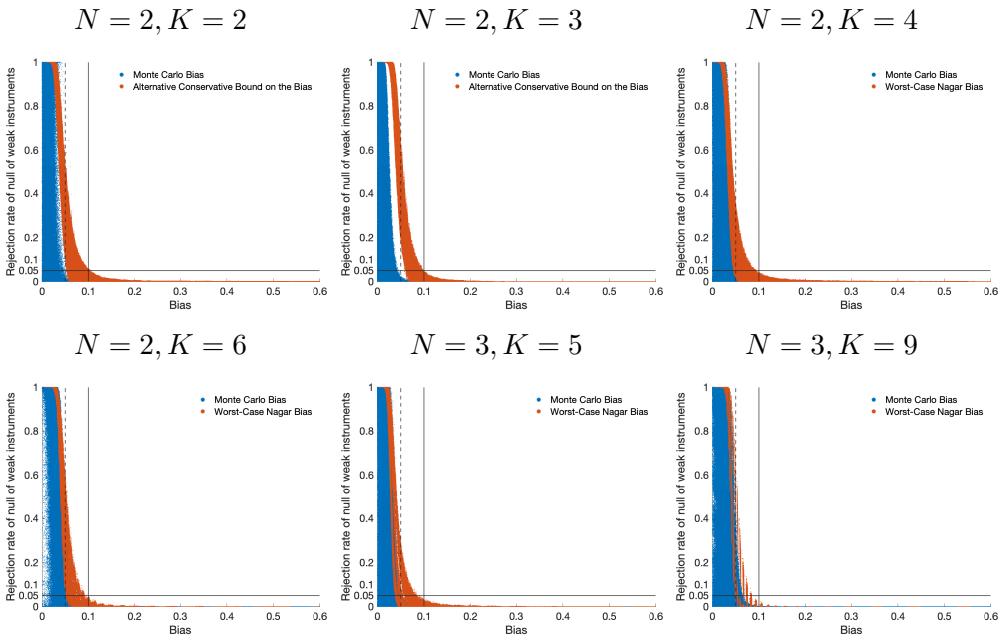


Notes: Rejection rates across 1,000 samples for each of five million DGPs generated as explained in the text for the test based on the absolute bias. The red dots show the rejection rates as a function of the worst-case Nagar bias or the alternative conservative bound on the bias. The blue dots show the rejection rates as a function of the Monte Carlo bias. For $N = 2, K = 2$, both consider the median bias. The vertical full line marks the bias tolerance level $\tau = 0.10$ in the null hypothesis, the dashed vertical line marks a bias level of 0.05 for reference, and the horizontal full line plots the nominal size $\alpha = 0.05$.

Figure G.3 shows that the empirical rejection rates never meaningfully exceed 0.05 at worst-case Nagar bias levels of 0.10 or higher. The rejection rates are frequently below 0.05. This is not surprising given our use of a bounding asymptotic distribution for the test statistic $g_{\min}$, which implies that the test is conservative by construction. As there are no DGPs with meaningful positive size distortions, the test controls size well at the nominal level. As discussed in the main text, the alternative conservative bound is effective in bounding the Monte Carlo bias in the simulations when $K \leq N + 1$. Section G.3 below will illustrate that this is not the case when, instead, the sharp bound on the Nagar bias is used as in the models with $K = N + 1$.

Despite the fact that the test is conservative, Figure G.3 shows that it nevertheless has meaningful power. The dashed vertical line marks a bias level of 0.05. At that level, the rejection rates in the $K = N + 2$ models, for example, rise as high as 0.434, while in the $K > N + 2$ models, the rejection rates reach up to 0.732. At lower – but still strictly positive – bias levels, the empirical rejection rates rise to unity for all DGPs. The simulations, therefore, demonstrate that our testing procedure is not excessively conservative.

FIGURE G.4: Size and Power of the First-Stage Test, Relative Bias



Notes: Rejection rates across 1,000 samples for each of five million DGPs generated as explained in the text for the test based on the relative bias. The red dots show the rejection rates as a function of the worst-case Nagar bias or the alternative conservative bound on the bias. The blue dots show the rejection rates as a function of the Monte Carlo bias. For $N = 2, K = 2$, both consider the median bias. The vertical full line marks the bias tolerance level $\tau = 0.10$ in the null hypothesis, the dashed vertical line marks a bias level of 0.05 for reference, and the horizontal full line plots the nominal size $\alpha = 0.05$.

The results for the relative bias test in Figure G.4 are very similar to those of the absolute bias test in Figure G.3. Just like the absolute bias test, the relative bias test controls size at the nominal level and has substantial power; for no DGP with Monte Carlo bias greater than $\tau = 0.10$ does the rejection rate exceed $\alpha = 0.05$. For worst-case relative bias of 0.05, the rejection rates in the $K = N + 2$ models, for example, rise as high as 0.376, while in the $K > N + 2$ models, the rejection rates reach up to 0.682. At lower – but still strictly positive – bias levels, the empirical rejection rates rise to unity for all DGPs.

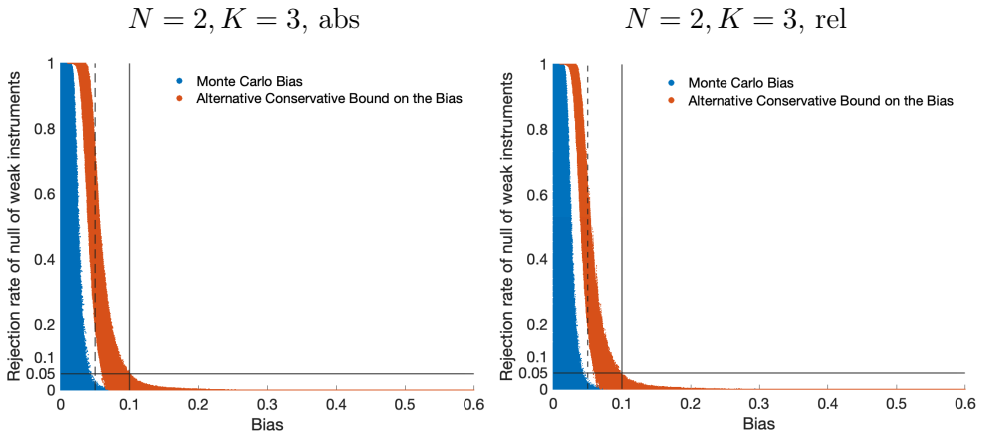
G.3 The Need for the Conservative Bound When $K = N + 1$.

The asymptotic simulations in the previous section demonstrate that our testing procedures perform as intended across a large number of randomly chosen DGPs. This is also true for models with $K \leq N + 1$ as long as the more conservative bound in part(ii) of Theorem 1 is used in those cases.

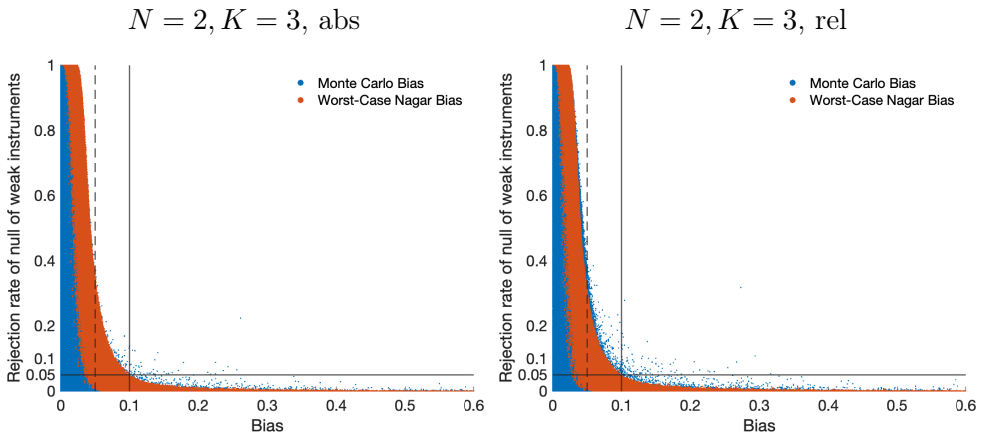
Figures G.5 and G.6 demonstrate the need for the more conservative bound in models with degrees of overidentification equal to one, $K = N + 1$. As

FIGURE G.5: Size and Power of the First-Stage Test, General Model

(a) Using More Conservative Bound on the Bias



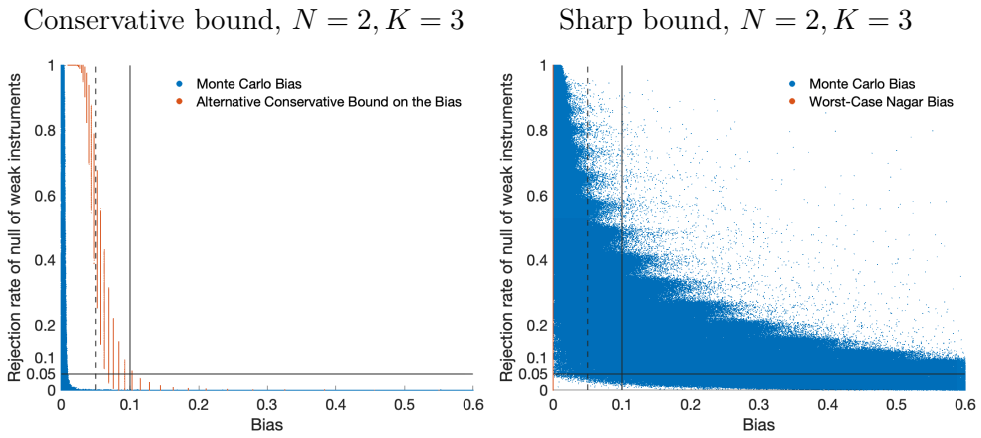
(b) Using Worst-Case Nagar Bias



Notes: Rejection rates across 1,000 samples for each of five million DGPs generated as explained in the text. The blue dots show the rejection rates as a function of the Monte Carlo bias. The red dots shows the rejection rates as a function of the alternative conservative bound on the bias (Panel a) or the worst-case Nagar bias (Panel b). The vertical full line marks the bias tolerance level $\tau = 0.10$, the null hypothesis, the dashed vertical line marks a bias level of 0.05 for reference, and the horizontal full line plots the nominal size $\alpha = 0.05$.

before, the various panels plot rejection frequencies against the Monte Carlo bias in blue. Panel (a) in Figure G.5 repeats the second panels in Figure G.3 and G.4 for ease of comparison and shows rejection frequencies based on the more conservative bound in red. In Panel (b), the rejection frequencies in red are instead based on the sharp bound on the Nagar bias (the worst-case Nagar bias), i.e. as in the models with $K > N + 1$. The results clearly illustrate that the worst-case Nagar bias is not adequate for a bias-based first-stage test in the models with $N = 2, K = 3$. There are a significant number of DGPs for which the Monte Carlo bias exceeds the bias tolerance while at the same time, the rejection rate is meaningfully above the nominal level of 0.05.

FIGURE G.6: Size and Power of the First-Stage Test, Homoskedastic Model



Notes: Rejection rates across 1,000 samples for each of five million DGPs generated as explained in the main text, except that $\mathbf{W}$ is of the Kronecker form. The blue dots show the rejection rates as a function of the Monte Carlo bias. The red dots shows the rejection rates as a function of the alternative conservative bound on the bias (left) or the worst-case Nagar bias (right). The absolute and relative bias tests are identical under homoskedasticity. The vertical full line marks the bias tolerance level $\tau = 0.10$ in the null hypothesis, the dashed vertical line marks a bias level of 0.05 for reference, and the horizontal full line plots the nominal size $\alpha = 0.05$.

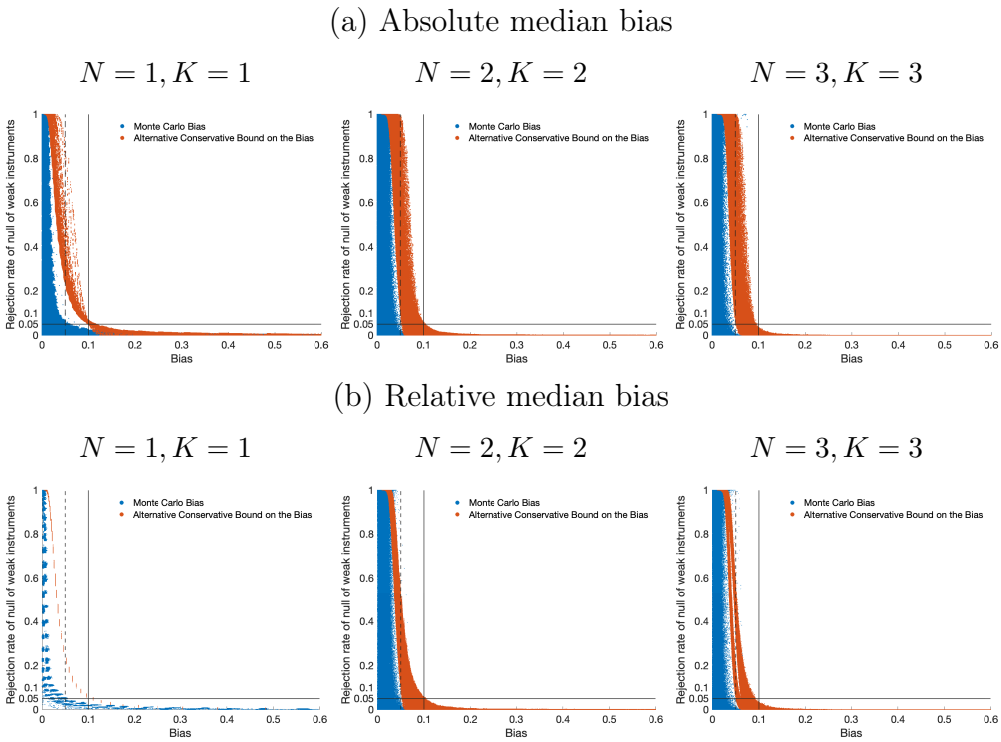
As explained in Section 2.5 of the main text, large positive size distortions can also occur in models with $K = N + 1$ under homoskedasticity and zero serial correlation. In that case, the worst-case Nagar bias is, in fact, analytically equal to zero. To illustrate the extent of the problem, Figure G.6 shows results from simulations for $N = 2$ and $K = 3$ where now $\mathbf{W}$ is restricted to have the exact Kronecker form in all five million DGPs. The left panel in Figure G.6 shows that the test based on the conservative bound continues to ensure that no rejection rates exceed 0.05 for DGPs with Monte Carlo bias larger than the tolerance level of 0.10. The right panel shows that, when the test is based on the worst-case Nagar bias, there is a very large number of DGPs for which rejection frequencies are well above the nominal size of 0.05 at bias levels exceeding 0.10.

G.4 Controlling Median Bias when $K = N$

As discussed in the main text, the bias criterion in Definition 2 does not exist for just-identified models, $K = N$, because the expected value $E[\beta^*_{2SLS}]$ does not exist. In order to maintain a bias-based testing approach also for just-identified models, we instead replace the expected value with the median of β^*_{2SLS} in the bias criterion.

When $K = N = 1$, the Nagar approximation can be easily adapted to

FIGURE G.7: Size and Power of the First-Stage Test, Median bias

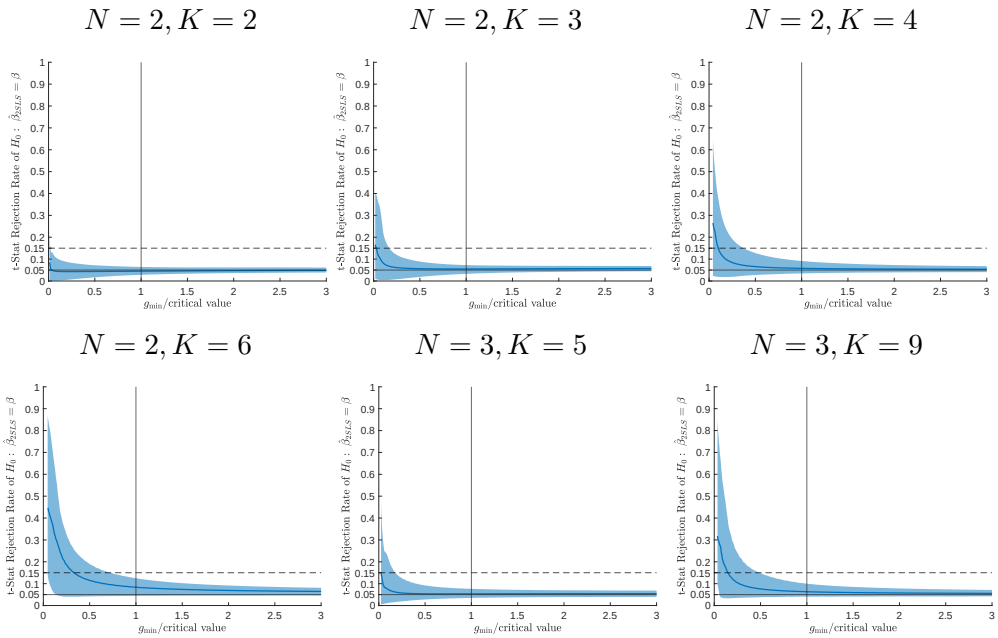


Notes: Rejection rates across 1,000 samples for each of five million DGPs generated as explained in the text. The blue dots show the rejection rates as a function of the Monte Carlo median bias. The red dots shows the rejection rates as a function of the alternative conservative bound on the bias, rescaled as described in the text in the case of $N = K = 1$. Panel (a) considers absolute bias and (b) relative bias. The vertical full line marks the bias tolerance level $\tau = 0.10$ in the null hypothesis, the dashed vertical line marks a bias level of 0.05 for reference, and the horizontal full line plots the nominal size $\alpha = 0.05$.

analytically approximate the median 2SLS bias, as described in Section 2.5. The first column of Figure G.7 plots the rejection rates from the resulting median bias-based tests across five million DGPs generated as described above. Now, the red region represents rejection rates plotted against the alternative conservative bound on the median bias criterion, which is the bound in part (ii) of Theorem 1 rescaled by median$(\chi^2_1) = 0.455$. The blue region plots rejection rates against the Monte Carlo median bias. As the figure shows, both versions of the test control size effectively. In the simulations, there is no DGP for which the Monte Carlo median bias is greater than $\tau = 0.10$ while at the same time the rejection rate exceeds $\alpha = 0.05$. The simulation evidence, therefore, supports the validity of our testing approach based on the median bias for models with $N = K = 1$.

When $K = N > 1$, there is, unfortunately, no straightforward analytical approximation to the median bias using the Nagar approximation as in the $N = K = 1$ model. However, the bound in part (ii) of Theorem 1 remains an

FIGURE G.8: Size of t-Statistic Inference on β, Absolute Bias



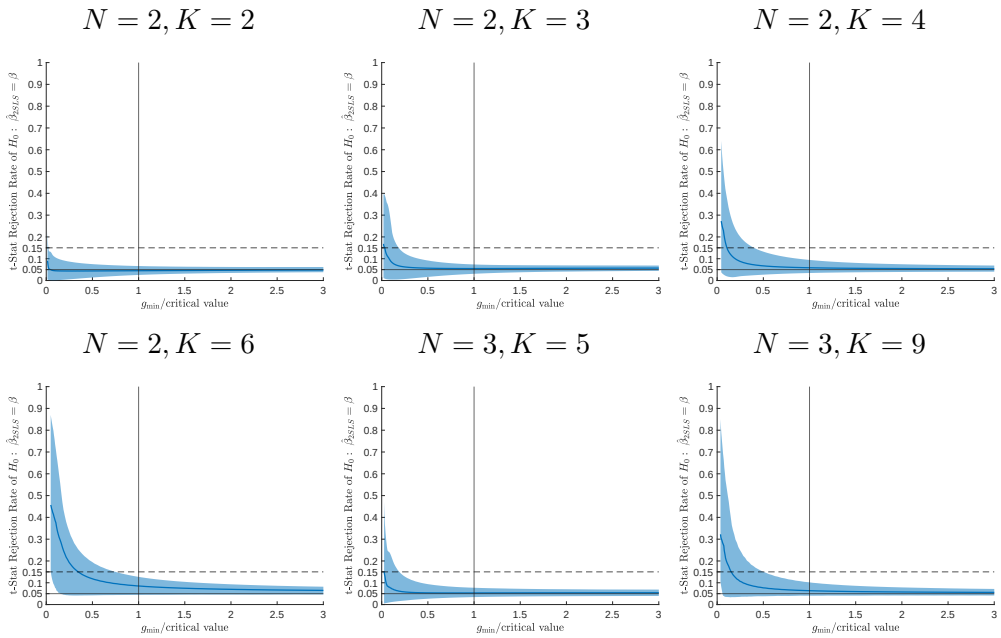
Notes: For each specification, we consider five million DGPs as described in the text. For each DGP, we take 1000 samples, and for each sample, we calculate the first-stage test statistic $g_{\min}$ and conduct a two-sided t-test for each element of β. The figure shows the average and 95 percentiles of the t-test rejection rates as a function of the average $g_{\min}$, relative to the absolute bias critical value, for 100 equally spaced bins.

effective, if more conservative, bound on the Monte Carlo median bias without modification. This can be seen in the middle and right columns of Figure G.7, which shows simulations for models with $N = K = 2$ and $N = K = 3$, respectively. The red dots plot rejection rates using the simplified critical values based on the conservative bound, while the blue dots plot rejection rates against the Monte Carlo median bias. There are no DGPs with Monte Carlo bias greater than 0.10 for which the rejection rate exceeds $\alpha = 0.05$.

G.5 Size Distortions of t-Statistic Inference on β

Alternative testing strategies for weak instruments can be based on controlling size distortions of Wald or t-statistic inference on β. The generalization of the size-based test of Stock and Yogo (2005) to heteroskedastic and serially correlated models is beyond the scope of this paper. Nevertheless, in this section, we explore the relationship between the test statistic $g_{\min}$ and the distortions of a standard two-sided t-test in Figures G.8 (absolute bias) and G.9 (relative bias). The t-tests are for the null hypothesis that a given element in $\hat{\beta}_{2SLS}$ equals the true value. Each panel shows binned averages of the rejection rates across the N t-tests in the five million DGPs as a function of

FIGURE G.9: Size of t-Statistic Inference on β, Relative Bias



Notes: For each specification, we consider five million DGPs as described in the text. For each DGP, we take 1000 samples, and for each sample, we calculate the first-stage test statistic $g_{\min}$ and conduct a two-sided t-test for each element of β. The figure shows the average and 95 percentiles of the t-test rejection rates as a function of the average $g_{\min}$, relative to the relative bias critical value, for 100 equally spaced bins.

the average ratio of $g_{\min}$ to the corresponding critical value of the first-stage test. The shaded area plots the 95 percent interval of the rejection rates within each bin. The full horizontal line shows the 0.05 nominal level of the t-test. For reference, the dashed horizontal line marks the 0.15 level, corresponding to a common tolerance level of 0.10 in size-based tests of weak instruments.

Figure G.8 shows that the size distortions generally grow larger as $g_{\min}$ becomes smaller relative to the critical value. In addition, the size distortions vanish, at least within the central 95% interval, as $g_{\min}$ grows larger. On average across the DGPs, the t-tests lead to over-rejection for low values of $g_{\min}$ relative to the critical value. The size distortions are relatively small in the $N = 2, K = 2$ model even when $g_{\min}$ is well below the critical value of the bias-based test. The size distortions become more severe at lower values of $g_{\min}$ as the degree of over-identification increases. The highest rejection rate observed for $g_{\min}$/critical value ≥ 1 is 0.125, for $N = 2, K = 6$. Overall, these patterns are qualitatively the same as those discussed in Stock and Yogo (2005) for CHSU models. They indicate that size distortions are well controlled (with a tolerance of 0.10) at values of $g_{\min}$ well below those required to control bias at $\tau = 0.10$ when the number of instruments is small. In

their results, as the number of instruments increases, $g_{\min}$ eventually needs to exceed the threshold for the bias-based test to control size in t-statistic inference on β; the same appears to be the case here if these simulations were extended to larger K. Figure G.9 repeats the exercise, plotting against the ratio of $g_{\min}$ to the relative bias critical values. The results are visually indistinguishable. The highest rejection rate observed for $g_{\min}$/critical value ≥ 1 is 0.126 for $N = 2, K = 6$.

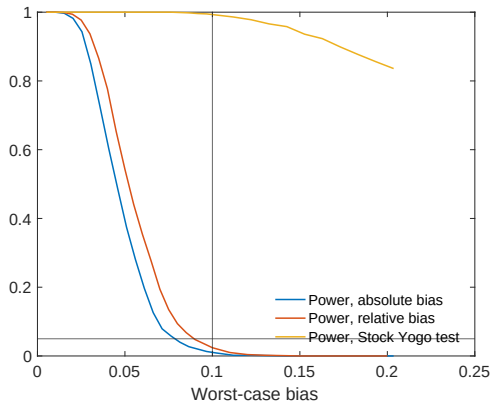
The relationships shown in Figures G.8 and G.9 naturally suggest an adjustment for t-based confidence intervals based on $g_{\min}$ similar to the one suggested recently in Lee et al. (2022) for $N = 1, K = 1$ models based on the first-stage F-statistic. We leave the development of such a procedure for future work.

H Empirical Simulations

This appendix presents simulation results based on a DGP that is calibrated to the empirical application of Ramey and Zubairy (2018) in Section 5. The purpose of these simulations is to evaluate the performance of the robust first-stage tests in an empirically realistic setting. More specifically, we construct a DGP that is calibrated to the regression associated with the first impulse response horizon for the specification that uses the full sample and the slack measure as the regime indicator, see Panel (a) in Figure 3. To construct this DGP, we first orthogonalize Z, w, v, to the controls (i.e. lagged variables). We take the point estimates of β and Π as given. We then use a kernel estimator (based on the Bartlett kernel) applied to the squared instruments and errors to estimate the time paths of their variances. After normalizing the estimated residuals by the corresponding time-varying volatilities, we separately estimate VAR models for Z and w, v to fit the serial correlation patterns present in the data. With these estimated processes in hand, we can then simulate data by drawing Gaussian innovations to the estimated VAR processes, computing the implied simulated paths for the normalized variables, and then multiplying them by the estimated empirical volatility paths. We scale the singular values of Π to match different values for $\lambda_{\min}$.

Figure H.10 plots power curves for various tests in this empirical design. As the power curve in yellow shows, the Stock-Yogo test exhibits very large size distortions, with an empirical size of essentially 1 when the worst-case bias is at the tolerance level of 0.10. In contrast, the robust absolute bias test (in blue) is conservative, with a null rejection rate of about 0.01, although power rises sharply to the left (0.38 for bias of 0.05). The robust relative bias

FIGURE H.10: Power of tests in empirical simulations



Notes: Power curves for various tests in simulations calibrated to an empirical specification from Ramey and Zubairy (2018), as described in the text. Blue is the absolute bias test, red is the relative bias test. The Stock and Yogo (2005) test is plotted in yellow against the worst-case Nagar bias under the absolute bias criterion.

test (in red) also effectively controls size: the null rejection rate is 0.02, and the test has meaningful power (0.54 at a relative bias of 0.05).

While we only report results calibrated to a single specification and impulse horizons, we found that the results for other specifications and horizons are qualitatively similar.

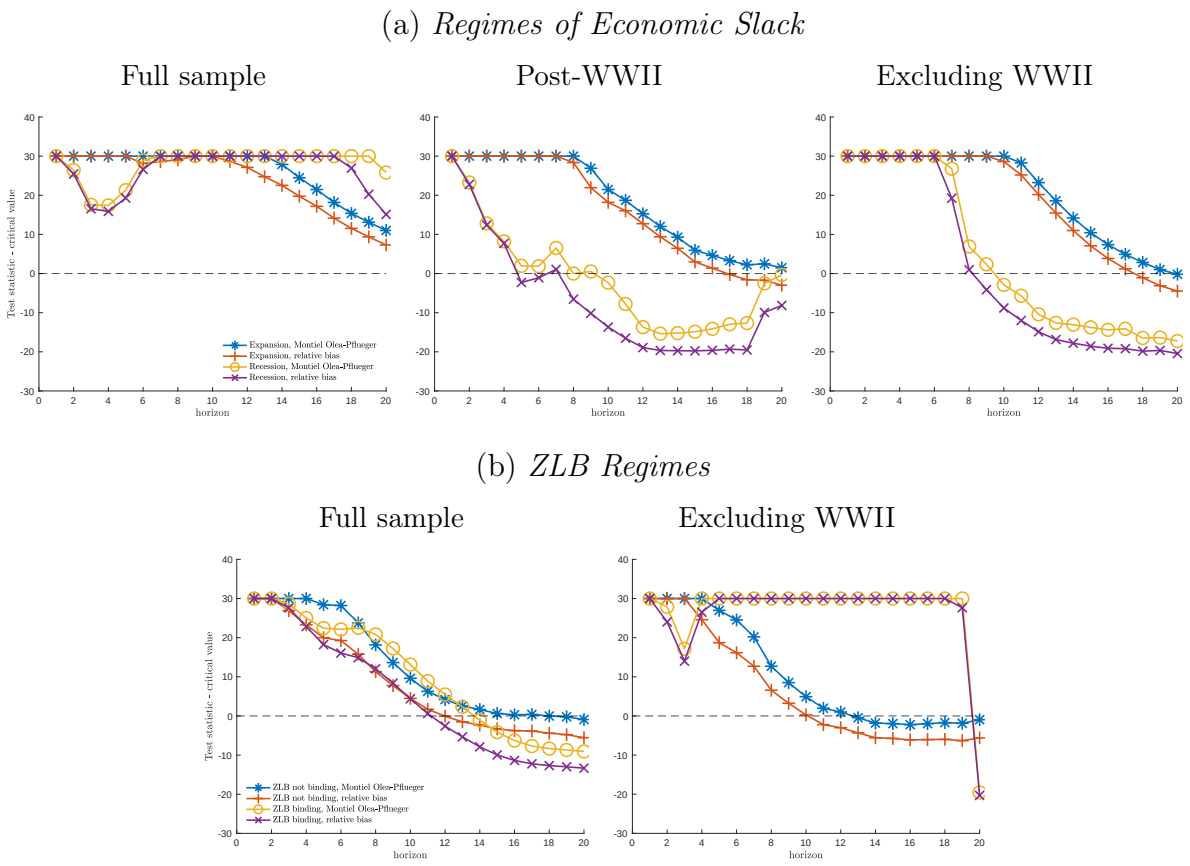
I Additional Results for the Empirical Application

Figure I.1 reports the first-stage test results for the regime subsamples, replicating those in the original Ramey and Zubairy (2018) paper. Within each regime subsample (expansion/recession, or in/out of ZLB), there is a just a single endogenous regressor, $N = 1$, and there are two instruments, $K = 2$. The model for the regime subsamples therefore has a degree of overidentification of one. As explained in the main text, this is one of the cases where the critical values from our relative bias test will not coincide with those from the Montiel Olea and Pflueger (2013) test. The figure therefore reports the results for both tests.

The blue and yellow lines in Figure I.1 report the difference between the effective F-statistic and the Montiel Olea and Pflueger (2013) critical values for $\tau = 0.10$ and $\alpha = 0.05$. As in Ramey and Zubairy (2018), we cap the results at 30 for visibility. These test results exactly replicate those reported in Figure 4 and 10 in Ramey and Zubairy (2018).

The red and purple lines in Figure I.1 report the differences between the effective F-statistic – which is identical to our $g_{\min}$ statistic when $N = 1$ –

FIGURE I.1: Robust Weak IV Test Results in Regime Subsample Regressions with $N = 1$ and $K = 2$



Notes: All panels report first-stage test results for regime subsamples within different sample periods: 1890-2015, 1947-2015 (post-WWII), and 1890-2015 excluding WWII. The blue line and yellow lines report the difference between the test statistics and Montiel Olea and Pflueger (2013) $\tau = 0.10$ and $\alpha = 0.05$ critical values. As in Ramey and Zubairy (2018), we cap the results at 30 for visibility. The red and purple lines show the corresponding results when using our relative bias critical values.

and the critical values from our relative bias test for $\tau = 0.10$ and $\alpha = 0.05$. As we opt for a more conservative bound on the bias in models that are over-identified with degree one, the resulting higher critical values lead to more failures of the first-stage test. Generally speaking, this means that the instruments are considered weak 1 to 4 horizons earlier than when using the Montiel Olea and Pflueger (2013) values.

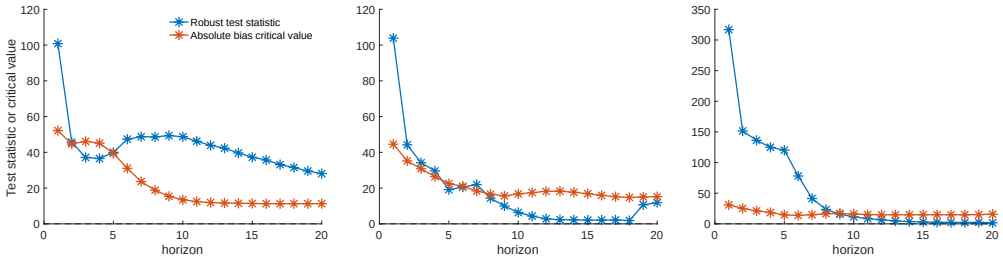
Figure 3 in the main text reports the difference between the test statistic and the critical values for the full model in (19) with two endogenous variables, $N = 2$, and four instruments, $K = 4$. Figure I.2 separately reports the values for the robust test statistic $g_{\min}$ and the absolute bias critical values; Figure I.3 does the same for the relative bias.

Finally, Figure I.4 reports results for the Andrews (2018) procedure for

FIGURE I.2: Robust Weak IV Test Statistics and Absolute Bias Critical Values

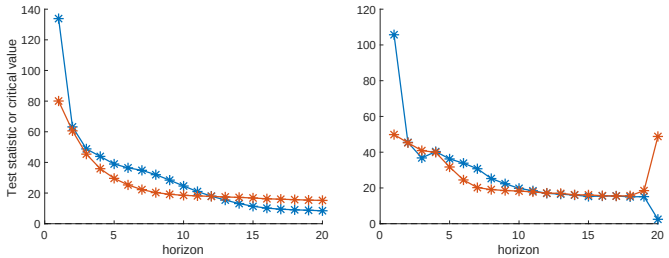
(a) *Government Spending Interacted with Indicator of Slack*

Full sample | Post-WWII | Excluding WWII



(b) *Government Spending Interacted with ZLB Indicator*

Full sample | Excluding WWII

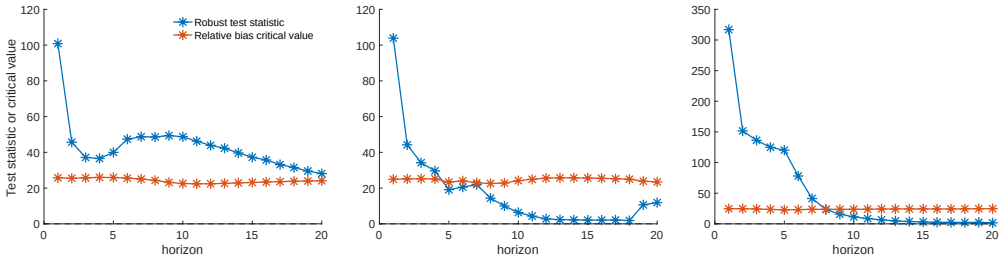


Notes: Panel (a) reports results for specifications with government spending interacted with an indicator for whether the economy was in a state of slack, using combined instruments for different sample periods: 1890-2015, 1947-2015 (post-WWII), and 1890-2015 excluding WWII. The blue line reports the robust test statistics for the interacted regression, and the red line reports the $\tau = 0.10$ and $\alpha = 0.05$ robust critical values based on the absolute bias criterion. Panel (b) reports analogous results for specifications with government spending interacted with an indicator for whether monetary policy is constrained by the zero lower bound for different sample periods, 1890-2015 and 1890-2015 excluding WWII.

FIGURE I.3: Robust Weak IV Test Statistics and Relative Bias Critical Values

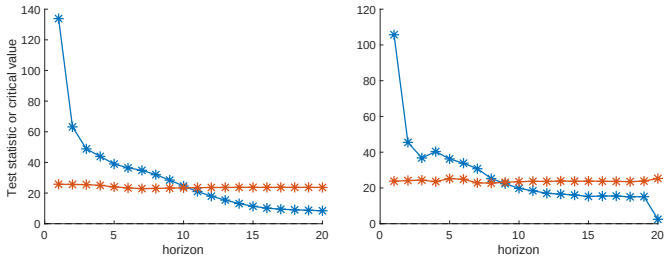
(a) *Government Spending Interacted with Indicator of Slack*

Full sample | Post-WWII | Excluding WWII



(b) *Government Spending Interacted with ZLB Indicator*

Full sample | Excluding WWII

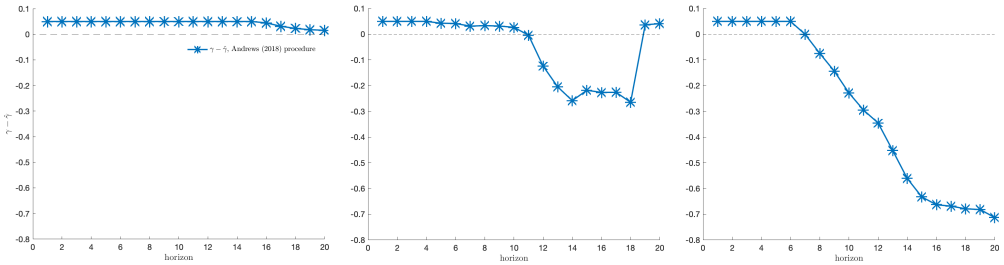


Notes: Panel (a) reports results for specifications with government spending interacted with an indicator for whether the economy was in a state of slack, using combined instruments for different sample periods: 1890-2015, 1947-2015 (post-WWII), and 1890-2015 excluding WWII. The blue line reports the robust test statistics for the interacted regression, and the red line reports the $\tau = 0.10$ and $\alpha = 0.05$ robust critical values based on the relative bias criterion. Panel (b) reports analogous results for specifications with government spending interacted with an indicator for whether monetary policy is constrained by the zero lower bound for different sample periods, 1890-2015 and 1890-2015 excluding WWII.

FIGURE I.4: Results for the Andrews (2018) procedure

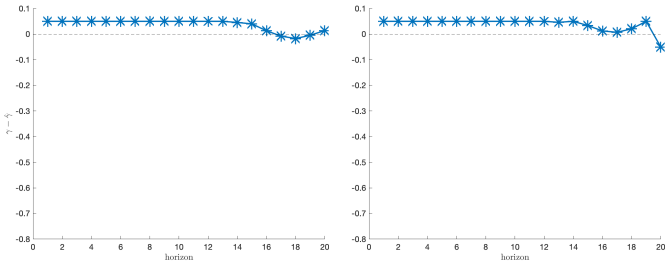
(a) *Government Spending Interacted with Indicator of Slack*

Full sample | Post-WWII | Excluding WWII



(b) *Government Spending Interacted with ZLB Indicator*

Full sample | Excluding WWII



Notes: Panel (a) reports results for specifications with government spending interacted with an indicator for whether the economy was in a state of slack, using combined instruments for different sample periods: 1890-2015, 1947-2015 (post-WWII), and 1890-2015 excluding WWII. The blue line reports the difference between a maximum coverage distortion of $\gamma = 0.10$ and $\hat{\gamma}$, the distortion cutoff supported by the data; negative values are evidence of weak instruments. Panel (b) reports analogous results for specifications with government spending interacted with an indicator for whether monetary policy is constrained by the zero lower bound for different sample periods, 1890-2015 and 1890-2015 excluding WWII.

detecting weak instruments. The blue lines plot the difference between a maximum coverage distortion of $\gamma = 0.10$ and $\hat{\gamma}$, the distortion cutoff supported by the data. Negative values are evidence of weak instruments; note that the values are capped above at 0.05 due to the calibration of $\gamma_{\min}$ in the Sun (2018) `twostepweakiv` Stata package. A value of zero indicates that non-robust (Wald) inference leads to a coverage distortion of $\gamma = 0.10$, the maximum distortion tolerated, and the cut-off for weak instruments. Note that this is akin to a size-based test, so the results are not expected to align with those for the bias-based tests reported above. This procedure indicates strong instruments for more specifications than the bias-based tests. However, the results are qualitatively similar, with instruments weaker at longer horizons. Moreover, omitting WWII or starting the sample after WWII generally reduces evidence of strong instruments, much as for the bias-based tests.

Online Appendix References

Alzer, Horst (1997). "On Some Inequalities for the Gamma and Psi Functions". In: *Mathematics of Computation* 66.217, pp. 373–389.

Andrews, Isaiah (2018). "Valid Two-Step Identification-Robust Confidence Sets for GMM". In: *The Review of Economics and Statistics* 100.2, pp. 337–348.

Imhof, J. P. (1961). "Computing the distribution of quadratic forms in normal variables". In: *Biometrika* 48.3-4, pp. 419–426.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter (2022). "Valid t-Ratio Inference for IV". In: *American Economic Review* 112.10, pp. 3260–90.

Montiel Olea, José Luis and Carolin Pflueger (2013). "A Robust Test for Weak Instruments". In: *Journal of Business & Economic Statistics* 31.3, pp. 358–369.

Ramey, Valerie A. and Sarah Zubairy (2018). "Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data". In: *Journal of Political Economy* 126.2, pp. 850–901.

Skeels, Christopher L. and Frank Windmeijer (2018). "On the Stock–Yogo Tables". In: *Econometrics* 6.4.

Stock, James and Motohiro Yogo (2005). "Testing for Weak Instruments in Linear IV Regression". In: *Identification and Inference for Econometric Models*. Ed. by Donald W.K. Andrews. New York: Cambridge University Press, pp. 80–108.

Sun, Liyang (2018). *TWOSTEPWEAKIV: Stata module to implement two-step weak-instrument-robust confidence sets for linear instrumental-variable (IV) models*. Statistical Software Components, Boston College Department of Economics.