A Robust Test for Weak Instruments for 2SLS with Multiple Endogenous Regressors*

Daniel J. Lewis
University College London

Karel Mertens†
Federal Reserve Bank of Dallas
CEPR

December 2, 2025

Abstract

We develop a test for instrument strength based on the bias of two-stage least squares (2SLS) that (1) generalizes the tests of Stock and Yogo (2005) and Sanderson and Windmeijer (2016) to be robust to heteroskedasticity and autocorrelation, and (2) extends the Montiel Olea and Pflueger (2013) robust test for models with a single endogenous regressor to multiple endogenous regressors. Our test can be based either on Stock and Yogo's (2005) absolute bias criterion or on the 2SLS bias relative to Montiel Olea and Pflueger's (2013) worst-case benchmark. We also develop extensions to test whether weak instruments cause bias in individual 2SLS coefficients. In simulations, our test controls size and is powerful, and we provide efficient code packages for its practical implementation. We demonstrate our testing procedures in the context of the estimation of state-dependent fiscal multipliers as in Ramey and Zubairy (2018).

JEL classification: C26, C36.
Keywords: Instrumental Variables, Weak Instruments Test, Multiple Endogenous Regressors, Heteroskedasticity, Serial Correlation.

*The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Banks of Dallas or the Federal Reserve System. We are grateful to Francesca Molinari as editor and four anonymous referees, to José Luis Montiel Olea and Carolin Pflueger for helpful discussions and for sharing their computer code, to Nikolay Gospodinov and Frank Windmeijer for great discussions, and to Isaiah Andrews, Lutz Kilian, Sarah Zubairy, Jonathan Shaw, and seminar/conference participants at Amsterdam, Bocconi, Cambridge, CEMFI, LSE, KU Leuven, Surrey, Dolomiti Macro Meetings 2024, Federal Reserve System Econometrics Meeting, Harvard Conference in Honor of Stock and Watson, IAAE 2023, and NBER SI EFFE 2023 for helpful comments. All errors are ours.

†*Contact*: mertens.karel@gmail.com, tel: +(214) 922-6000, fax: +(214) 922-6302

Empirical researchers using instrumental variables (IV) estimation frequently report first-stage F-statistics – or with multiple endogenous regressors, Cragg and Donald (1993) statistics – to assess instrument relevance. These statistics are typically compared to critical values tabulated by Stock and Yogo (2005), which are based on the maximum bias of two-stage least-squares (2SLS) that can occur when the first-stage coefficients are statistically close to zero. This "bias-based" testing procedure, however, requires the assumption of conditionally homoskedastic and serially uncorrelated (CHSU) errors and is generally invalid if that assumption does not hold. As discussed in a recent survey by Andrews et al. (2019), in practice, this often means that researchers assume CHSU errors for the purpose of a first-stage test but make different assumptions when conducting inference for the 2SLS estimates.

In an important paper, Montiel Olea and Pflueger (2013) introduce a new statistic – the "effective F-statistic" – for models with general non-CHSU errors. Their test is based on a second-order Taylor approximation of the 2SLS bias relative to a worst-case benchmark, and the limiting distribution and associated critical values depend on the application-specific structure of the robust covariance matrix of the reduced-form and first-stage parameters. A key limitation, however, is that this test only applies to models with a single endogenous regressor. Andrews et al. (2019) point to the lack of a robust weak instruments test for models with multiple endogenous regressors as an important remaining gap in the literature.[1]

The contribution of this paper is to fill that gap by providing a robust weak instrument test that allows for an arbitrary number of endogenous regressors and, in doing so, generalizes Stock and Yogo's (2005) bias-based test to be heteroskedasticity-and-autocorrelation robust. When applied to a transformed regression, our test also accommodates the alternative setting of Sanderson and Windmeijer (2016) in which instruments are weak because the first-stage coefficient matrix is near to having a rank reduction of one rather than near rank zero as in the setting of Stock and Yogo (2005). Like that in Stock and Yogo (2005), the bias criterion for our test is a weighted quadratic loss in the asymptotic bias of the 2SLS estimates. The bias criterion evaluates either the (standardized) absolute bias as defined by Stock and Yogo (2005), which, as in their case, can also be interpreted as the 2SLS bias relative to the

[1]Researchers sometimes report the robust F-statistic, or the Kleibergen and Paap (2006) statistic, along with Stock and Yogo (2005) critical values. In models with a *single* endogenous regressor, both statistics coincide with the effective F-statistic of Montiel Olea and Pflueger (2013) when there is a single instrument, while Windmeijer (2025) shows that the robust F-statistic is also informative for the bias of an overidentified GMM estimator with a specific weighting matrix. However, like Andrews et al. (2019), we are unaware of a theoretical justification for using robust F- or Kleibergen and Paap (2006) statistics to assess instrument strength in models with *multiple* endogenous regressors.

maximum possible OLS bias, or the bias relative to a worst-case benchmark for the 2SLS bias established by Montiel Olea and Pflueger (2013). As a further contribution, we show how to conduct robust weak instrument tests for both bias criteria when only the 2SLS coefficient on one of several endogenous regressors is of interest. We restrict attention to 2SLS as it is by far the most commonly used IV estimator in practice.

A key challenge for a practical robust weak instruments test for models of arbitrary dimensions is that the critical values depend on the covariance matrix of the first-stage and reduced-form coefficients. Since this matrix is application-specific, the critical values cannot be tabulated in advance as for CHSU models. To develop a test that is computationally feasible, we follow Montiel Olea and Pflueger (2013) and characterize the weak instrument set in terms of the "Nagar bias", a second-order Taylor approximation to the bias following Nagar (1959). We show analytically that the Nagar bias has an upper bound that is inversely proportional to the minimum eigenvalue of the concentration parameter but depends otherwise only on covariances that can be consistently estimated using heteroskedasticity-and-autocorrelation- or cluster-robust methods. The minimum eigenvalue of the concentration parameter is, therefore, the object of the test, as it suffices to characterize the upper bound on the Nagar bias. The null hypothesis is that the minimum eigenvalue of the concentration parameter, which quantifies the strength of the instruments, is in the set of values for which the maximum Nagar bias is greater than a tolerance level τ.

We propose a test statistic that generalizes the Cragg-Donald statistic adopted by Stock and Yogo (2005) and also nests the effective F-statistic of Montiel Olea and Pflueger (2013) for models with a single endogenous regressor. The statistic's asymptotic distribution is that of the minimum eigenvalue of a matrix with elements that are traces of block partitions of a real non-central Wishart random matrix, which, as a result, is intractable. Just as Stock and Yogo (2005), we therefore make use of an approximate limiting distribution. Specifically, we show that the second and third cumulants of our generalized test statistic are bounded by expressions that depend only on consistently estimable covariances and the minimum eigenvalue of the concentration parameter. The right tail of the limiting distribution of the test statistic can typically be approximated simply by matching the bounding cumulants using an Imhof (1961) distribution.[2] The accompanying code computes the upper bound on the Nagar bias efficiently, leading to trivial computation times in most applications. For unusually large-dimensional models,

[2]Our accompanying code ensures an appropriate distribution even when this is not the case.

we also provide a simplified – but more conservative – bound on the Nagar bias that avoids numerical optimization entirely. In addition, we show that simple adjustments to the tolerance level suffice to test for weak instruments bias in individual coefficients and that applying the testing procedures to a transformed regression implements a robust version of the test in Sanderson and Windmeijer (2016).

We assess the performance of the test through both finite-sample and asymptotic simulations. In a finite-sample design adapted from Sanderson and Windmeijer (2016), our test demonstrates good size and power properties, despite relying on critical values derived from an approximate limiting distribution. The test exhibits sharp power increases at bias levels well above zero and not too far below the chosen tolerance level. Asymptotic simulations confirm the accurate size and good power of the test across millions of randomly generated DGPs with varying numbers of endogenous regressors and instruments. They also allow us to verify the accuracy of the Nagar approximations in bounding the 2SLS bias.[3] Additional finite-sample simulations calibrated to our empirical application also demonstrate accurate size and attractive power.

As an empirical application of our test, we consider the Ramey and Zubairy (2018) estimates of state-dependent government spending multipliers. Starting from existing empirical specifications, the authors introduce a second endogenous regressor by interacting government spending with an indicator for the state of the business cycle or for the monetary policy regime. This application is an example of how multiple endogenous regressors often arise in practice, as it is common for researchers to explore specifications with interactions between an endogenous regressor and other variables. We find that our robust test regularly leads to different conclusions regarding instrument strength than the Stock and Yogo (2005) test. When results of the Montiel Olea and Pflueger (2013) test in the separate regime subsamples are in conflict, the outcome of our generalized test for the full sample is dictated by the more weakly identified regime. Multiple endogenous regressors, of course, arise in many other contexts as well, including in time series, cross-sectional, and panel data models. Our generalized test for weak instruments should, therefore, be useful for a broad range of applications.

Besides the aforementioned papers, there exist several other approaches to evaluating instrument strength. In addition to their bias-based test, Stock

[3]We focus on the median 2SLS bias in just-identified models since the mean 2SLS bias does not exist in those models. We find that the Nagar approximation can perform poorly when the degree of overidentification is less than two. We therefore propose different bounds in those cases, which we also recommend for the Montiel Olea and Pflueger (2013) test when there is a single endogenous regressor.

and Yogo (2005) propose a test based on the size (the rejection rate under the null hypothesis) of Wald inference for CHSU models. The test detects whether instruments are weak enough to lead to size distortions (over-rejection) above a specified tolerance. Mikusheva and Sun (2021) develop a size-based test that is robust to heteroskedasticity and potentially many instruments, but not autocorrelation and is limited to models with one endogenous regressor. While developing a generalized size-based test is beyond the scope of this paper, we explore the relationship between our test statistic and the size distortion in second-stage t-statistic inference in the asymptotic simulations. Andrews (2018) offers a distinctly different diagnostic for identification strength that imposes no restrictions on the error covariances and is broadly applicable to GMM problems. In this approach, both identification-robust and non-robust confidence sets are computed for the parameters. The relationship between the sets, which is linked to the size distortion of standard inference, determines whether identification is strong enough to proceed with inference methods that are not robust to weak instruments. Finally, Carrasco and Doukali (2021) introduce a bias-based first-stage test that is robust to heteroskedasticity and many instruments. However, as the Montiel Olea and Pflueger (2013) test, it only applies to models with a single endogenous regressor.

1 Model and Summary of Testing Procedures

This section provides a non-technical summary of the model assumptions and testing procedures.

Model Our test applies to linear instrumental variables models with N endogenous regressors and $K \geq N$ instruments,

$$y = Y\beta + u, \tag{1}$$
$$Y = Z\Pi + v, \tag{2}$$

where $\beta \in \mathbb{R}^N$ contains the parameters of interest and $\Pi \in \mathbb{R}^{K\times N}$ contains the first-stage parameters. The econometrician observes $y \in \mathbb{R}^T$, $Y \in \mathbb{R}^{T\times N}$, and $Z \in \mathbb{R}^{T\times K}$, where T is the sample size and individual observations are indexed by t. Without loss of generality, we assume that there are no additional exogenous regressors and that $Z'Z/T = I_K$.[4]

The reduced form of (1) is $y = Z\Pi\beta + w$, where $w = v\beta + u$. Assume that $T^{-\frac{1}{2}}[Z'w \;\; \text{vec}(Z'v)']' \xrightarrow{d} \mathcal{N}(0, \mathbf{W})$, where $\text{vec}(\cdot)$ is the vectorization operator.

[4]In the presence of additional exogenous regressors, y, Y, and Z can simply be replaced by the residuals of those variables after projection on those regressors.

Given $Z'Z/T = I_K$, this is simply the standard high-level assumption that the reduced-form and first-stage OLS estimates are asymptotically normal. The weak instruments test of Stock and Yogo (2005) requires the asymptotic covariance matrix of the reduced-form and first-stage OLS estimates, $\mathbf{W}$, to be of the Kronecker product form $\Sigma_{wv} \otimes I_K$, where Σ_{wv} is the covariance matrix of $[w_t,\ v_t']'$ and the covariance of Z is standardized to I_K. The Kronecker structure for $\mathbf{W}$ generally arises only under the CHSU assumption. The test in this paper lets $\mathbf{W}$ be any positive definite matrix without requiring that $N = 1$ as in Montiel Olea and Pflueger (2013). In other words, our test can be used in models with an arbitrary number of endogenous regressors and arbitrary patterns of heteroskedasticity, serial correlation, or clustering, as long as a consistent estimate of $\mathbf{W}$ is available.

Testing Procedure and Interpretation Our robust test for weak instruments assesses the null hypothesis that the bias of the 2SLS estimator $\hat{\beta}_{2SLS}$ exceeds some tolerance level τ, with significance level less than or equal to α. The baseline test assumes that this bias arises because all elements of the matrix of first-stage parameters Π are statistically near zero. The magnitude of the 2SLS bias depends on $\mathbf{W}$, which can be consistently estimated by the econometrician, but also on β and Π, which cannot be consistently estimated if the instruments are weak. To ensure validity regardless of the values of β and Π, the null hypothesis addresses the bias associated with the most adverse configuration of those unknown parameters.

To obtain a criterion that aggregates the bias across the multiple elements of $\hat{\beta}_{2SLS}$, we follow Stock and Yogo (2005) and adopt a weighted quadratic loss function divided by a scalar that determines the criterion's quantitative interpretation. We consider two options for the weighting and scaling, both of which ensure that the bias criterion is invariant to arbitrary full-rank linear combinations of the model variables. The first criterion uses the same weights and scaling as the absolute bias criterion of Stock and Yogo (2005). The weights in this criterion standardize the endogenous regressors such that they have unit standard deviations and are orthogonal, and the scaling is in units of σ_u, the standard deviation of the model error u_t. The resulting criterion can be interpreted as the absolute bias in standardized units of $\sigma_u^{-1}\Sigma_Y^{\frac{1}{2}}$ where Σ_Y is the asymptotic covariance of Y. As in Stock and Yogo (2005), the absolute bias criterion also has the interpretation as the bias relative to the maximum OLS bias after standardizing the endogenous regressors.[5]

[5]The criterion differs subtly from the relative bias criterion also discussed in Stock and Yogo (2005), which expresses the 2SLS bias relative to the OLS bias rather than to the *maximum* OLS bias. In CHSU

The criterion, therefore, helps separate the problems of endogeneity and weak instruments and allows the econometrician to interpret the units of the bias using expert judgment on the largest possible magnitude of the OLS bias. The second bias criterion, which we label the "relative bias" criterion, is a direct extension of the criterion proposed by Montiel Olea and Pflueger (2013). This criterion uses weights and scaling such that, in a worst-case scenario for an ad-hoc approximation to the 2SLS bias, the bias criterion equals unity. The relative bias is, therefore, interpretable in terms of a percentage of this worst-case benchmark. The relative bias criterion is identical to that of Montiel Olea and Pflueger (2013) when $N = 1$, and also coincides with the absolute bias criterion in the special case of CHSU models.

In practice, our test for weak instruments consists of the following steps:

1. Replace all variables with the residuals from regressions on a constant and any other included exogenous regressors (Frisch-Waugh-Lovell). Standardize the instruments such that $Z'Z/T = I_K$.
2. Regress y on Z (reduced form) and Y on Z (first stage). Obtain the robust covariance estimator $\hat{\mathbf{W}}$ of choice (e.g., heteroskedasticity-robust, heteroskedasticity-and-autocorrelation robust, or clustered) for the reduced-form and first-stage estimates. For the absolute bias criterion, also obtain an estimate $\hat{\Sigma}_{wv}$ of the covariance matrix of the reduced-form and first-stage errors.
3. Compute the test statistic, $g_{\min}$,

$$g_{\min} = T^{-1}\,\text{mineval}\{\hat{\Phi}^{-\frac{1}{2}}Y'ZZ'Y\hat{\Phi}^{-\frac{1}{2}}\}, \tag{3}$$

where

$$\hat{\Phi} = \begin{bmatrix} \text{Tr}(\hat{\mathbf{W}}_{2,11}) & \text{Tr}(\hat{\mathbf{W}}_{2,12}) & \dots & \text{Tr}(\hat{\mathbf{W}}_{2,1N}) \\ \vdots & \ddots & & \vdots \\ \text{Tr}(\hat{\mathbf{W}}_{2,N1}) & \dots & \dots & \text{Tr}(\hat{\mathbf{W}}_{2,NN}) \end{bmatrix}. \tag{4}$$

$\hat{\mathbf{W}}_2$ is the estimated covariance matrix of the first-stage coefficients (the lower $NK \times NK$ diagonal block of $\hat{\mathbf{W}}$), $\hat{\mathbf{W}}_{2,ij}$ is the block corresponding

settings, the distinction is irrelevant since both criteria are identical under the most adverse choice of β. In non-CHSU settings, however, the ratio of the 2SLS bias to the OLS bias is not a feasible criterion as it can be made arbitrarily large for a suitable choice of β. While the interpretation of the absolute criterion as a percentage of the maximum OLS bias is ultimately the same as that of the Stock and Yogo (2005) relative criterion under CHSU, we maintain their "absolute bias" terminology to make a clear distinction with the criterion of Montiel Olea and Pflueger (2013).

to the covariance of coefficients from the Y_i equation with those from the Y_j equation, and mineval$\{\cdot\}$ is the smallest eigenvalue.

4. Given a choice of bias criterion (absolute or relative), a user-supplied bias threshold τ (e.g., 0.10) and a desired significance level α (e.g., 0.05), compare $g_{\min}$ to the application-specific critical value. The latter depends on all blocks of $\hat{\mathbf{W}}$ and, under the absolute bias criterion, also on $\hat{\Sigma}_{wv}$. If $g_{\min}$ exceeds the critical value, reject the null hypothesis of weak instruments.

We provide a Matlab file, `gweakivtest.m`, to implement steps 1-4 using y, Y, Z, and X as inputs, with several standard options for obtaining $\hat{\mathbf{W}}$. The Matlab function `gweakivtest_critical_values.m` calculates the critical value in step 4 given values for τ and α, a choice of bias criterion, and the estimates $\hat{\mathbf{W}}$ and $\hat{\Sigma}_{wv}$ (if needed). The computation of the critical values is very fast, even in models with fairly large dimensions. For very large-dimensional models, the code package also provides the option to calculate alternative critical values that are essentially instantaneous to compute, although these will be more conservative.

Relationship with Existing Tests The test statistic $g_{\min}$ nests existing test statistics as follows: When $N = 1$, $g_{\min}$ equals the effective F-statistic of Montiel Olea and Pflueger (2013). When $N = K = 1$, $g_{\min}$ equals the robust F-statistic, and when $N = 1$ and the CHSU assumption is imposed, it equals the regular F-statistic. When $N > 1$ and the CHSU assumption is imposed, $g_{\min}$ equals the Cragg and Donald (1993) statistic.

As our test relies on the same second-order approximation of the 2SLS bias, for models with $N = 1$, the critical values under the relative bias criterion are essentially identical to those of Montiel Olea and Pflueger (2013). However, there are two exceptions. First, in just-identified models, the mean bias of 2SLS does not exist. Hence, whenever $N = K$, we prefer to base the test on the median bias of 2SLS instead. When $N = K = 1$, this means that the critical value for our relative bias test differs from that of Montiel Olea and Pflueger (2013), and will be smaller for the same τ and α. Second, in models that are overidentified of degree one, we show – both analytically and in simulations – that the second-order approximation does not provide a reliable bound on the 2SLS bias. We therefore recommend a more conservative bound for those models. For $N = 1$, $K = 2$, this means that our test generally applies larger critical values than Montiel Olea and Pflueger (2013). Under the absolute bias criterion, our test provides an alternative to the Montiel

Olea and Pflueger (2013) test for models with $N = 1$ and any K, in case researchers find the absolute bias criterion more appealing.

Under the CHSU assumption, the absolute and relative versions of the test are identical and essentially the same as the Stock and Yogo (2005) test. The critical values are, in that case, very similar to theirs, though they are not identical since our test uses a second-order approximation rather than Monte Carlo simulations to evaluate the bias. The other difference is that, unlike Stock and Yogo (2005), our test covers models with $K \leq N + 1$. We extend the range of models by considering the median bias when $K = N$, and by adopting a conservative but numerically reliable bound on the 2SLS bias when $K = N + 1$.

Extensions to Other Hypothesis Tests The test described above is concerned with the 2SLS bias in all N coefficients that arises when the rank of Π is statistically close to zero. In practice, researchers may sometimes be concerned with the bias in just a single coefficient. We show that the same testing procedures and code can be applied in that case with only a simple adjustment to τ, which is developed in Section 3.

Weak instrument bias can occur even if the rank of Π is not statistically close to zero. For example, instrument relevance can also fail when one of the columns of Π is statistically close to a linear combination of the other columns, such that Π is not near rank zero but near a column rank reduction of one. We show how our test can also be used in this setting by applying it to a transformed regression that has a single endogenous regressor. The accompanying program, `gweakivtest_LRR1.m`, conducts all required steps automatically. The program allows for testing for bias in all N coefficients in this setting, thereby generalizing the test described Sanderson and Windmeijer (2016). The program also allows testing for bias in the individual 2SLS estimates in this alternative setting. Section 3 describes in more detail how to conduct these additional tests for weak instruments in practice.

Proper Use of the Test Weak instruments generally have two consequences: bias in the point estimates and invalidity of standard (i.e. t-statistic) inference. Our testing procedures are informative about the former but do not address the latter. Weak-instrument-robust inference methods are necessary to do so. As best practice for empirical research, we recommend reporting the results of our generalized weak instruments test alongside the point estimate(s) and weak-instrument-robust confidence intervals or sets using, e.g., the Anderson and Rubin (1949) statistic (see Andrews et al. (2019) for further

discussion). Recent contributions by Lee et al. (2022) and Keane and Neal (2022) have highlighted the pre-testing problem when conducting a first-stage F-test for weak instruments and then using t-statistic inference if weak instruments are rejected. Specifically, even if the instruments are found to be strong, the correlation between the first-stage F and t-statistics can induce severe size distortions. While first-stage tests cannot justify the use of standard inference methods without accounting for this correlation, they are nevertheless of independent interest since they contain important information about the bias in the 2SLS point estimates. Finally, we caution that our testing procedures can be performed under the assumption that either the rank of Π is close to zero as in Stock and Yogo (2005), or alternatively that the rank of Π is close to $N-1$, as in Sanderson and Windmeijer (2016). Hence, applied users should think carefully about the multivariate relationship between instruments and endogenous variables to determine the appropriate framework and test.

2 Testing the Null Hypothesis of Weak Instruments

This section presents the key results underlying our proposed test. Before doing so, we establish some specific notation: $||U||_2$ is the spectral norm of U (the positive square root of the maximum eigenvalue of UU', also the ℓ_2-norm if U is a vector), $\mathbb{P}^n$ is the set of positive definite $n \times n$ matrices, $\mathbb{O}^{n\times m}$ is the set of $n \times m$ orthonormal real matrices U such that $UU' = I_n$, $\mathcal{K}_{n,m}$ denotes the $n \times m$ commutation matrix such that $\mathcal{K}_{n,m}\,\text{vec}(U) = \text{vec}(U')$ where $U \in \mathbb{R}^{n\times m}$. We also define the special matrix $R_{n,m} = I_n \otimes \text{vec}(I_m)$. The dimension of $R_{n,m}$ is $nm^2 \times n$. For $U \in \mathbb{R}^{nm\times nm}$, the (i,j)-th element of $V = R'_{n,m}(U \otimes I_m)R_{n,m} \in \mathbb{R}^{n\times n}$ is $\text{Tr}(U_{ij})$ where $U_{ij} \in \mathbb{R}^{m\times m}$ is (i,j)-th block of U and $\text{Tr}(\cdot)$ is the trace (see (4) for an explicit example of such a matrix). For $U \in \mathbb{R}^{nm\times m}$, the i-th element of $V = R'_{n,m}\,\text{vec}(U') \in \mathbb{R}^n$ is equal to $\text{Tr}(U_i)$ where $U_i \in \mathbb{R}^{m\times m}$ is the i-th row block of U. Note that $R'_{n,m}R_{n,m} = mI_N$.

2.1 Local-to-Zero Asymptotic Representation of 2SLS

The 2SLS estimator for the model in (1)–(2) is

$$\hat{\beta}_{2SLS} \;=\; (Y'P_ZY)^{-1}Y'P_Zy, \tag{5}$$

where $P_Z = ZZ'/T$ is the projection matrix given the normalization $Z'Z/T = I_K$. Following Staiger and Stock (1997) and subsequent literature, we model weak instruments by assuming the first-stage relationship is local-to-zero,

Assumption 1. $\Pi = C/\sqrt{T}$, *where* $C \in \mathbb{R}^{K\times N}$ *is a fixed full-rank matrix.*

While weak instruments are a finite sample problem, the local-to-zero assumption is a modeling device to capture the limited informativeness of the instruments in an asymptotic framework. Specifically, Assumption 1 implies that all instruments are uniformly weak and that the OLS estimate of any element of Π has a nondegenerate limiting distribution even as T grows large. The Stock and Yogo (2005) test is also based on Assumption 1, but there exist other approaches for modeling weak identification in models with multiple endogenous regressors. We maintain Assumption 1 here and show later how our test can be used in alternative weak instruments settings.

The next set of assumptions allows us to characterize the asymptotic distribution of the 2SLS estimator under the local-to-zero assumption.

Assumption 2. *The following limits hold as* $T \to \infty$*:*

$$T^{-\frac{1}{2}} \begin{bmatrix} Z'w \\ \operatorname{vec}(Z'v) \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, \mathbf{W}), \tag{2.a}$$

$$\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}, \tag{2.b}$$

$$\hat{\Sigma}_{wv} \xrightarrow{p} \Sigma_{wv} = E\left[\begin{bmatrix} w_t \\ v_t \end{bmatrix}\begin{bmatrix} w_t \\ v_t \end{bmatrix}'\right], \tag{2.c}$$

where $\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_{12} \\ \mathbf{W}_{12}' & \mathbf{W}_2 \end{bmatrix} \in \mathbb{P}^{(N+1)K}$ *and* $\Sigma_{wv} = \begin{bmatrix} \sigma_w^2 & \sigma_{wv} \\ \sigma_{wv}' & \Sigma_v \end{bmatrix} \in \mathbb{P}^{N+1}$.

In part (a), a central limit theorem ensures that the first-stage and reduced-form OLS estimates are asymptotically normal with covariance $\mathbf{W}$. Part (b) stipulates the availability of a consistent estimator for $\mathbf{W}$, denoted as $\hat{\mathbf{W}}$. Likewise, part (c) specifies the availability of a consistent estimator $\hat{\Sigma}_{wv}$ for the covariance of the reduced-form and first-stage errors, Σ_{wv}. These high-level assumptions are standard, but researchers must, of course, confirm within their specific application whether assuming consistent estimation of $\mathbf{W}$ and Σ_{wv} is reasonable.[6]

The 2SLS estimator in (5) can be written as

$$\hat{\beta}_{2SLS} = \left(R_{N,K}'(s_{ZY}s_{ZY}' \otimes I_K)R_{N,K}\right)^{-1} R_{N,K}' \operatorname{vec}(s_{Zy}s_{ZY}'), \tag{6}$$

where $s_{Zy} = T^{-\frac{1}{2}}Z'y$ and $s_{ZY} = T^{-\frac{1}{2}}\operatorname{vec}(Z'Y)$. This more complicated expression reformulates the 2SLS estimator in terms of random vectors whose

[6]The assumption that $\mathbf{W}$ is consistently estimable can be questionable, for example, in the presence of many instruments, see Mikusheva and Sun (2021).

asymptotic distributions are given in Assumption 2. Specifically, define the random vectors $\eta_1 \in \mathbb{R}^K$, $\eta_2 \in \mathbb{R}^{NK\times 1}$,

$$\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{0}_K\\ \text{vec}(C)\end{pmatrix}, \mathbf{S}\right) \tag{7}$$

where $\mathbf{S} \in \mathbb{P}^{(N+1)K}$ is the asymptotic covariance of $T^{-\frac{1}{2}}[Z'u \;\; \text{vec}(Z'v)']'$, and is partitioned as $\mathbf{W}$ with

$$\begin{aligned}\mathbf{S}_1 &= \mathbf{W}_1 + (\beta' \otimes I_K)\mathbf{W}_2(\beta \otimes I_K) - (\beta' \otimes I_K)\mathbf{W}_{12}' - \mathbf{W}_{12}(\beta \otimes I_K),\\ \mathbf{S}_{12} &= \mathbf{W}_{12} - (\beta' \otimes I_K)\mathbf{W}_2 \;,\; \mathbf{S}_2 = \mathbf{W}_2.\end{aligned} \tag{8}$$

Proposition 1 characterizes the distribution of $\beta^*_{2SLS} = \hat{\beta}_{2SLS} - \beta$.

Proposition 1. *Under Assumptions 1 and 2,* $s_{ZY} \xrightarrow{d} \eta_2$ *and* $s_{Zy} \xrightarrow{d} (\beta' \otimes I_K)\eta_2 + \eta_1$*, and thus*

$$\hat{\beta}_{2SLS} - \beta \xrightarrow{d} \beta^*_{2SLS} = \left(R_{N,K}'(\eta_2\eta_2' \otimes I_K)R_{N,K}\right)^{-1} R_{N,K}' \,\text{vec}(\eta_1\eta_2').$$

Proof. The proposition follows directly from the stated assumptions, the expression for $\hat{\beta}_{2SLS}$ in (6), and the continuous mapping theorem. □

Under the local-to-zero assumption, β^*_{2SLS} converges to a quotient of quadratic forms in normal random variables and $\hat{\beta}_{2SLS}$ is therefore not a consistent estimator of β. The asymptotic bias of the 2SLS estimator is the expected value $E[\beta^*_{2SLS}]$, which has no known analytical form except for the special case in which $\mathbf{W} = \Sigma_{wv} \otimes I_K$, $N = 1$, and $K > 1$, see Skeels and Windmeijer (2018). Note that when $K = N$ – and depending on assumptions – also when $K = N+1$, the expected value $E[\beta^*_{2SLS}]$ generally does not exist.[7]

Finally, we define the concentration parameter for models with general $\mathbf{W}$ and $N \geq 1$. The concentration parameter is the (standardized) mean of the "denominator" of the 2SLS estimator and measures the signal-to-noise ratio in the first stage. Since it is the center of the distribution of the 2SLS denominator, it is critical for characterizing the weak instruments distribution of $\hat{\beta}_{2SLS}$, and thus its asymptotic bias.

Definition 1. *The concentration parameter for (2) is*

$$\Lambda = \Phi^{-\frac{1}{2}} C'C \Phi^{-\frac{1}{2}} \;,$$

[7] See, for example, Basmann (1961), Mariano (1972), Kinal (1980), Phillips (1980), or Skeels and Windmeijer (2018).

where $\Phi = R'_{N,K}(\mathbf{S}_2 \otimes I_K)R_{N,K}$.

Under Assumptions 1 & 2, $E[R'_{N,K}(\eta_2\eta_2' \otimes I_K)R_{N,K}] = C'C + R'_{N,K}(\mathbf{S}_2 \otimes I_K)R_{N,K} = C'C + \Phi$, so the standardized non-centrality of the denominator of the 2SLS estimator is given by Λ. The $N \times N$ matrix Φ contains the traces of the $K \times K$ partitions of the $NK \times NK$ asymptotic covariance of the first-stage coefficients, $\mathbf{S}_2 = \mathbf{W}_2$. Here, it standardizes the concentration parameter to be unitless and invariant to full-rank linear combinations of regressors and instruments. The concentration parameter in Definition 1 nests special cases as follows: when $\mathbf{W}$ has the Kronecker form, $\mathbf{S}_2 = \Sigma_v \otimes I_K$ and $\Phi = K\Sigma_v$ such that $\Lambda = K^{-1}\Sigma_v^{-\frac{1}{2}}C'C\Sigma_v^{-\frac{1}{2}}$ which is the concentration parameter in Stock and Yogo (2005). For general $\mathbf{W}$ and $N = 1$, $\Lambda = ||C||_2^2 \operatorname{Tr}(\mathbf{W}_2)^{-1}$, which is the concentration parameter in Montiel Olea and Pflueger (2013).

2.2 Definition of Weak Instruments

As in Stock and Yogo (2005) and Montiel Olea and Pflueger (2013), the subject of our null hypothesis is a criterion for the bias of 2SLS. Specifically, we consider instruments weak when a weighted quadratic loss function of the asymptotic bias $E[\beta^*_{2SLS}]$ is large, either in an absolute or relative sense.

Definition 2. *The bias criterion for* $i \in \{abs,\ rel\}$ *is*

$$B_i = \sqrt{\frac{E\left[\beta^*_{2SLS}\right]' \Xi_i E\left[\beta^*_{2SLS}\right]}{b_i}} = ||E\left[\beta^*_{2SLS}\right]' \Xi_i^{\frac{1}{2}}||_2/\sqrt{b_i}$$

where $b_{abs} = \sigma_u^2$, $\Xi_{abs} = \Sigma_v$; $b_{rel} = \operatorname{Tr}\mathbf{S}_1$, $\Xi_{rel} = \Phi$.

The bias criterion aggregates the N elements of the bias $E[\beta^*_{2SLS}]$ through a quadratic loss function with weighting matrix Ξ_i, such that $B_i \geq 0$ and larger biases are penalized more heavily. The choice of the weighting matrix Ξ_i and the scalar b_i determines whether the bias criterion is expressed in absolute or relative terms. As mentioned earlier, later we will also consider tests that instead place all weight on a single element of $E\left[\beta^*_{2SLS}\right]$.

The absolute bias criterion, B_{abs}, uses the same weighting ($\Xi_{abs} = \Sigma_v$) and scaling ($b_{abs} = \sigma_u^2$) as Stock and Yogo's (2005) absolute bias criterion. Under Assumption 1, $Y'Y/T \xrightarrow{p} \Sigma_Y = \Sigma_v$, such that $||E\left[\beta^*_{2SLS}\right]' \Sigma_v^{\frac{1}{2}}||_2$ is the aggregated asymptotic bias in $\hat{\beta}_{2SLS}$ after standardizing the endogenous regressors to be orthogonal and have unit standard deviation. Because this bias metric still depends on the units of the dependent variable, it is further divided by the standard deviation of the error term, σ_u. As in Stock and Yogo (2005), the criterion can be interpreted as the bias of the 2SLS estimator in the

standardized units of $\sigma_u^{-1}\Sigma_Y^{\frac{1}{2}}$. The standardization ensures that the criterion is invariant to taking arbitrary full-rank linear combinations of Y or Z, or to re-scaling y. Since σ_u is also the upper bound on the norm of the asymptotic OLS bias weighted by $\Sigma_Y^{\frac{1}{2}}$, a value of unity means that the 2SLS bias is equal to the maximum bias OLS could attain.[8] As mentioned earlier, the units of the absolute bias criterion can thus also be interpreted as percentages of the largest possible OLS bias across all values of the structural coefficients.

The relative bias criterion in Definition 2, B_{rel}, instead extends the bias criterion used by Montiel Olea and Pflueger (2013) in their test. Under this relative criterion, the bias is expressed as a percentage of a "worst-case benchmark" for the 2SLS bias. As in Montiel Olea and Pflueger (2013), the magnitude of this worst-case benchmark follows from an ad-hoc approximation of $E[\beta^*_{2SLS}]$ in terms of a ratio of expectations:

$$E[\beta^*_{2SLS}] \approx \text{Tr}(\mathbf{S}_1)^{\frac{1}{2}}\Phi^{-\frac{1}{2}}(I_N + \Lambda)^{-1}\frac{\Phi^{-\frac{1}{2}}R'_{N,K}\,\text{vec}(\mathbf{S}_{12})}{\text{Tr}(\mathbf{S}_1)^{\frac{1}{2}}} \,. \tag{9}$$

Using this approximation, B_{rel} reaches a maximum of unity when the instruments are completely uninformative, so the concentration parameter Λ is zero, and when the error term u is a perfect linear combination of the first-stage errors, v, such that the fraction in (9) is an $N \times 1$ unit vector. The weighting by $\Xi_{rel} = \Phi$ and scaling by $\text{Tr}(\mathbf{S}_1)$ also imply that the relative bias criterion is invariant to taking full-rank linear combinations of Y or Z, or re-scaling y.

When $N = 1$ the relative bias becomes $B_{rel} = E\left[\beta^*_{2SLS}\right]\sqrt{\text{Tr}(\mathbf{S}_2)}/\sqrt{\text{Tr}(\mathbf{S}_1)}$ which is identical to that in Montiel Olea and Pflueger (2013) after replacing $E\left[\beta^*_{2SLS}\right]$ with a Nagar (1959) approximation. The relative bias criterion in Definition 2 is, therefore, a direct generalization of Montiel Olea and Pflueger (2013) to models with an arbitrary number of endogenous regressors. Note that, in CHSU models, $\text{Tr}(\mathbf{S}_1) = K\sigma_u^2$ and $\Phi = K\Sigma_v$, so the absolute and relative bias criterion in Definition 2 are identical in that case.

Given a choice of bias criterion, we define the weak instrument set as the set of models, over values for β and the first-stage parameters C, such that the Nagar approximation of the bias ($B_{i,n}$, described below) exceeds a tolerance level τ.

Definition 3. *The weak instrument set for $i \in \{abs,\ rel\}$ is*

$$\mathbb{B}_{i,\tau}(\mathbf{W}) = \{C \in \mathbb{R}^{N\times K}, \beta \in \mathbb{R}^N : B_{i,n} > \tau\} \,.$$

[8] Evaluating the absolute bias of OLS yields $\sigma_u^{-1}||\Sigma_v^{-\frac{1}{2}}[\sigma'_{wv} : \Sigma_v]\tilde{\beta}||_2$ with $\sigma_u = ||\Sigma_{wv}^{\frac{1}{2}}\tilde{\beta}||_2$ and $\tilde{\beta} = [1 : -\beta']'$. This is maximized at the largest singular value of $\Sigma_v^{-\frac{1}{2}}[\sigma'_{wv} : \Sigma_v]\Sigma_{wv}^{-\frac{1}{2}}$, which is unity.

The weak instrument set depends on $\mathbf{W}$, which can be consistently estimated by assumption, but also on the NK parameters in C and the N unknown parameters in β. Under the absolute bias criterion, $i = abs$, the weak instrument set also depends on Σ_{wv}, but we henceforth suppress this dependence for compactness. The objective of our test is to determine if the model in question is a member of the weak instrument set: $B_{i,n} > \tau$. Since we cannot estimate C and β consistently, the null hypothesis will be that an upper bound on $B_{i,n}$, as a function of $\mathbf{W}$, is greater than τ. While we define weak instruments in terms of the Nagar bias, we demonstrate via extensive simulations (Appendix **??**) that our test also performs well with respect to the true bias since the upper bounds on the Nagar bias generally provide reliable upper bounds on the true bias (computed through numerical integration).

2.3 Characterizing the Boundary of the Weak Instrument Set

A practical generalized test for weak instruments requires a scalar statistic that is informative for whether C lies within the weak instrument set for any possible value of β. Moreover, the boundary of the set over the remaining parameters must be relatively easy to compute since – unlike in the CHSU model – this boundary depends on $\mathbf{W}$ (and on Σ_{wv} if $i = abs$) and thus differs for each application. To address these challenges, we follow Montiel Olea and Pflueger (2013) and adopt an analytical approximation based on Nagar (1959), which appears in the definition of the weak instruments set above.

We start with the following lemma presenting a useful decomposition of the bias criterion in Definition 2.

Lemma 1. *Under Assumptions 1 & 2, the bias criterion in Definition 2 can be decomposed as* $B_i = ||\tilde{\Xi}_i^{\frac{1}{2}} \mathbf{h} \rho_i||_2$, *where*

$$\mathbf{h} = KE\left[\left(R'_{N,K}(\mathcal{S}(l+\psi)(l+\psi)'\mathcal{S}' \otimes I_K)R_{N,K}\right)^{-1} R'_{N,K}\left(\mathcal{S}(l+\psi)\psi'\mathcal{S}^{-1} \otimes I_K\right)\right],$$
$$\rho_i = \left(\Phi^{-\frac{1}{2}} \otimes I_{K^2}\right) \mathrm{vec}\left(\mathbf{S}_{12}\right) / \sqrt{b_i}\ ,$$

$\tilde{\Xi}_i = \Phi^{-\frac{1}{2}} \Xi_i \Phi^{-\frac{1}{2}}$, $l = \mathbf{S}_2^{-\frac{1}{2}} \mathrm{vec}(C')$, $\psi = \mathbf{S}_2^{-\frac{1}{2}}(\eta_2 - \mathrm{vec}(C')) \sim \mathcal{N}(0, I_{NK})$, *and* $\mathcal{S} = ((\Phi/K)^{-\frac{1}{2}} \otimes I_K)\mathbf{S}_2^{\frac{1}{2}}$.

Proof. See Online Appendix **??**. □

Lemma 1 formulates the bias as the ℓ_2-norm of the product of the $N \times N$ matrix $\tilde{\Xi}_i$, the $N \times NK^2$ matrix $\mathbf{h}$ and the $NK \times 1$ vector ρ_i. The matrix $\mathbf{h}$ is the expected value of a random matrix that is a function of ψ, a vector of i.i.d standard normal variables. This expected value – when it exists – also

depends on location parameters C and on $\mathbf{W}_2$. The vector ρ_i depends on $\mathbf{W}$ and β and the choice for the scaling factor b_i, and $\tilde{\Xi}_i$ depends on $\mathbf{W}_2$ and the weighting matrix Ξ_i.

In general, there is no tractable analytical expression for the integral underlying the expectation in $\mathbf{h}$, which is required to evaluate the bias. Whereas Stock and Yogo (2005) evaluate this integral using Monte Carlo methods, we adopt a Nagar (1959) approximation to $\mathbf{h}$, which we denote by $\mathbf{h}_n$. This analytical approximation avoids costly numerical approximation over a generally high-dimensional parameter grid each time critical values are needed. The approximation leads to the following definition.

Definition 4. *The Nagar bias is* $B_{i,n} = ||\tilde{\Xi}_i^{\frac{1}{2}} \mathbf{h}_n \rho_i||_2$, *where* $\mathbf{h}_n$ *is a Nagar (1959) approximation of* $\mathbf{h}$ *around* $\psi = 0$.

The Nagar approximation $\mathbf{h}_n$ is derived analytically in Appendix A. The Nagar bias $B_{i,n}$ still depends on C, β, and $\mathbf{W}$ (and Σ_{wv} for $i = abs$). The following theorem shows that upper bounds on $B_{i,n}$ can be indexed to a scalar unknown, the minimum eigenvalue of the concentration parameter, $\lambda_{\min}$.

Theorem 1. *The Nagar bias has the following bounds for* $i = abs, rel$*:*

(i) $B_{i,n} \leq B^*_{i,n}(\mathbf{W}, \lambda_{\min}) = \lambda_{\min}^{-1} \mathcal{B}_i(\mathbf{W})$, $\mathcal{B}_i(\mathbf{W}) = K^{-\frac{1}{2}} ||\tilde{\Xi}_i^{\frac{1}{2}}||_2 \sup_{L_0 \in \mathbb{O}^{N \times K}} \{||M_1 (I_N \otimes L_0 \otimes L_0) M_2 \Psi_i||_2\}$,

(ii) $B^*_{i,n}(\mathbf{W}, \lambda_{\min}) \leq \lambda_{\min}^{-1} \mathcal{B}^s_i(\mathbf{W})$, $\mathcal{B}^s_i(\mathbf{W}) = ||\tilde{\Xi}_i^{\frac{1}{2}}||_2 \min\{(2(N+1)/K)^{\frac{1}{2}} ||M_2 \Psi_i||_2, ||\Psi_i||_2\}$,

where $\lambda_{\min} = \text{mineval}\{\Lambda\}$, *and*

$$\Psi_{abs} = (\mathcal{S}\mathbf{W}_2^{-\frac{1}{2}} [\mathbf{W}_{12} : \mathbf{W}_2]' \otimes I_K) R_{N+1,K} \Sigma_{wv}^{-\frac{1}{2}},$$
$$\Psi_{rel} = (\mathcal{S}\mathbf{W}_2^{-\frac{1}{2}} [\mathbf{W}_{12} : \mathbf{W}_2]' \otimes I_K) R_{N+1,K} (R'_{N+1,K} (\mathbf{W} \otimes I_K) R_{N+1,K})^{-\frac{1}{2}},$$
$$M_1 = R'_{N,N} \left(I_{N^3} + (\mathcal{K}_{N,N} \otimes I_N)\right), \; M_2 = R_{N,K} R'_{N,K}/(N+1) - I_{NK^2} \; .$$

Proof. See Appendix A. □

Part (i) of the theorem characterizes an upper bound on the Nagar bias, $B^*_{i,n}(\mathbf{W}, \lambda_{\min})$, that only depends on $\mathbf{W}$ (and Σ_{wv} if $i = abs$) and $\lambda_{\min}$, the minimum eigenvalue of the concentration parameter in Definition 1. Importantly, $B^*_{i,n}(\mathbf{W}, \lambda_{\min})$ is often a sharp upper bound on the Nagar bias, by which we mean that it is attained for at least some value of β and eigenstructure of Λ. More specifically, the upper bound is always sharp under the relative bias criterion since, in that case, $\tilde{\Xi}_i = I_N$. The bound is also sharp when $\mathbf{W}$ has the Kronecker form regardless of the bias criterion chosen, since $\tilde{\Xi}_i = I_N$ in

that case as well. For general $\mathbf{W}$, the bound is sharp under the absolute bias criterion when $N = 1$, but it is not necessarily sharp when $N > 1$. We will refer to the upper bound $B^*_{i,n}(\mathbf{W}, \lambda_{\min})$ as the "worst-case Nagar bias". The $\mathcal{B}_i(\mathbf{W})$ term in the expression for the worst-case Nagar bias is proportional to the largest singular value of a matrix that depends on the $NK-(N+1)/2$ nuisance parameters in the orthonormal matrix L_0, which needs to be obtained numerically. Part (ii) of the theorem provides an alternative simplified upper bound that requires no numerical optimization. This alternative bound is, in general, not sharp and (weakly) larger than the worst-case Nagar bias, but it is easy to obtain numerically even for very large-dimensional models.

Theorem 1 addresses two key challenges for a test of weak instruments in models with multiple endogenous variables. The first is statistical. Theorem 1 provides a straightforward mapping between the upper bound on the Nagar bias and a scalar parameter, $\lambda_{\min}$: $B^*_{i,n}(\mathbf{W}, \lambda_{\min}) = \lambda^{-1}_{\min}\mathcal{B}_i(\mathbf{W})$. The proof of the theorem shows that the Nagar bias under the most adverse scenario for β is non-increasing in all eigenvalues of the concentration parameter, such that the worst-case Nagar bias occurs when all eigenvalues of Λ equal $\lambda_{\min}$. The Nagar approximation, therefore, provides an analytical justification for $\lambda_{\min}$ as the object of the test, whereas the parallel justification in Stock and Yogo (2005) is based on simulation evidence or the many-weak instruments limit.

Second, more practically, the optimization problem underlying $\mathcal{B}_i(\mathbf{W})$ is relatively straightforward. The general problem of optimizing $B_{i,n}$ over β and C subject to mineval$\{\Lambda\} = \lambda_{\min}$ is problematic because of the large dimension and the presence of many local maxima. The proof in Appendix A shows that optimizing over β reduces to a straightforward maximum eigenvalue problem for either choice of bias criterion. Moreover, when all eigenvalues of Λ are equal to $\lambda_{\min}$, the bias depends additionally only on L_0. Optimization over L_0 has a smaller dimension and, importantly, can exploit numerical algorithms specialized for orthonormal matrices. In the code accompanying this paper, we use the curvilinear search algorithm of Wen and Yin (2013), which leads to trivial computation times even for relatively large N and K. When N and/or K are so large that the optimization becomes prohibitive, the more conservative simplified bound in part (ii) of Theorem 1 can be used instead. Finally, because the bounds are inversely proportional to the minimum eigenvalue of the concentration parameter, the threshold value of that minimum eigenvalue for a given bias tolerance level τ is given simply by $\mathcal{B}_i(\mathbf{W})/\tau$ (or $\mathcal{B}^s_i(\mathbf{W})/\tau$), and requires no additional root-finding operation.

The computational advantages of the Nagar approximation come at the cost of a potential loss in accuracy relative to using Monte Carlo integra-

tion to evaluate $\mathbf{h}$, as in Stock and Yogo (2005). Monte Carlo integration, however, cannot exploit the same simplifications in the optimization over the nuisance parameters and requires an additional root-finding operation to find the threshold value of the minimum eigenvalue. Because of the dependence of $\mathcal{B}_i(\mathbf{W})$ on $\mathbf{W}$, this threshold value must be computed separately for each application and cannot be tabulated in advance as a function of only K and N, unlike when $\mathbf{W}$ has the Kronecker form. In Online Appendix **??**, we compare the Nagar bias to the bias computed using Monte Carlo integration across millions of DGPs and for various values of N and K. The main finding is that the Nagar bias is generally close to the bias obtained via numerical integration when the degree of overidentification is greater than one, $K > N+1$. Importantly, the Nagar approximation is especially accurate in the most relevant range for the bias tolerance in practice, that is, in a neighborhood of 0.10.[9] Models with $K \leq N+1$ are more complicated and are discussed separately in Section 2.5.

2.4 Test Statistic and Critical Values

A test of the null hypothesis of weak instruments can be based on a test of whether the minimum eigenvalue of Λ is smaller or equal to a threshold value $\lambda^*_{\min}(\tau)$. More formally, the null and alternative hypotheses for the test are

$$H_0 : \lambda_{\min} \leq \lambda^*_{\min,i}(\tau) \quad \text{vs.} \quad H_1 : \lambda_{\min} > \lambda^*_{\min,i}(\tau), \tag{10}$$

where $\lambda^*_{\min,i}(\tau) = \mathcal{B}_i(\mathbf{W})/\tau$ (or $\mathcal{B}^s_i(\mathbf{W})/\tau$ in the case of the simplified test). The null hypothesis is that the minimum eigenvalue of the concentration parameter does not exceed the threshold beyond which the selected upper bound on the bias becomes smaller than the tolerance level τ. By Theorem 1, this guarantees that $B_{i,n} \leq \tau$. Hence, although the null hypothesis is in terms of $\lambda_{\min}$, the test is informative for $B_{i,n}$, the object of interest in Definition 3. Under the alternative, the minimum eigenvalue is larger than this threshold, such that the selected upper bound on the Nagar bias is below the tolerance level. A test that controls size for $\lambda_{\min}$ will also control size for $B_{i,n}$.

The following proposition presents our statistic to test the null hypothesis.

Proposition 2. *Define the test statistic*

$$g_{\min} = \text{mineval}\{\hat{\Phi}^{-\frac{1}{2}}(Y'P_ZY)\hat{\Phi}^{-\frac{1}{2}}\},$$

[9] The simulation results in Online Appendix **??** also show that the relative bias is bounded above by unity under numerical integration, but not when evaluated using the Nagar approximation, except when $N = 1$ as shown in Theorem 1.3 of Montiel Olea and Pflueger (2013).

where $\hat{\Phi} = R'_{N,K}(\hat{\mathbf{W}}_2 \otimes I_K)R_{N,K}$. *Then, under Assumptions 1 and 2,*

$$g_{\min} \xrightarrow{d} \text{mineval}\{R'_{N,K}(\zeta \otimes I_K)R_{N,K}/K\},$$

where the $NK \times NK$ *random matrix* $\zeta = \mathcal{S}(l+\psi)(l+\psi)'\mathcal{S}'$ *has a noncentral Wishart distribution,* $\zeta \sim \mathcal{W}(1, \Sigma, \Omega)$*, with* 1 *degree of freedom, scale matrix* $\Sigma = \mathcal{S}\mathcal{S}' \in \mathbb{P}^{NK}$*, and a matrix of noncentrality parameters* $\Omega = \Sigma^{-1}\mathcal{S}ll'\mathcal{S}'$.[10]

Proof. The proposition follows from Slutsky's theorem, the continuous mapping theorem, and $Y'P_ZY \xrightarrow{d} R'_{N,K}\left(\mathbf{S}_2^{\frac{1}{2}}(l+\psi)(l+\psi)'\mathbf{S}_2^{\frac{1}{2}'} \otimes I_K\right)R_{N,K}$, which implies the stated distribution of ζ. □

The test statistic $g_{\min}$ is a generalization of the Cragg and Donald (1993) statistic, rescaled appropriately to account for heteroskedasticity and autocorrelation. The matrix $\hat{\Phi}$ contains the block-wise traces of the estimated covariance of the first-stage coefficients, $\hat{\mathbf{W}}_2$. The scaling by $\hat{\Phi}^{-\frac{1}{2}}$ is precisely the one required so that there is a direct mapping between the non-centrality matrix of ζ and Λ. In the special case of the CHSU model, $\hat{\mathbf{W}}_2 = \hat{\Sigma}_v \otimes I_K$, such that $\hat{\Phi} = K\hat{\Sigma}_v$, and $g_{\min} = K^{-1}\,\text{mineval}\{(\hat{\Sigma}_v)^{-\frac{1}{2}}Y'P_ZY(\hat{\Sigma}_v)^{-\frac{1}{2}}\}$, which is the Cragg and Donald (1993) statistic. When $N = 1$, $\hat{\Phi} = \text{Tr}(\hat{\mathbf{W}}_2)$, such that $g_{\min} = Y'P_ZY/\,\text{Tr}(\hat{\mathbf{W}}_2)$, which is the effective F-statistic of Montiel Olea and Pflueger (2013).

Define $F^{-1}_{C,\mathbf{W}_2}(\alpha)$ as the upper α-quantile of the asymptotic distribution of $g_{\min}$, and let

$$c\,(\alpha, \mathbf{W}_2, x) = \sup_{C \in \mathbb{R}^{N\times K},\ \beta \in \mathbb{R}^N} F^{-1}_{C,\mathbf{W}_2}(\alpha)\mathbb{1}_{\lambda_{\min}<x}, \tag{11}$$

where $\mathbb{1}_{\lambda_{\min}<x}$ is an indicator function. Using this critical value, we characterise the asymptotic size of the test.

Proposition 3. *Under Assumptions 1-2 the test is pointwise asymptotically valid, that is*

$$\sup_{\mathbb{B}_{i,\tau}(\mathbf{W})} \lim_{T\to\infty} \text{Prob}\left(g_{\min} > c\left(\alpha, \hat{\mathbf{W}}_2, \mathcal{B}_i(\hat{\mathbf{W}})/\tau\right)\right) \leq \alpha.$$

Proof. See Appendix **??**. □

Intuitively, setting $x = \mathcal{B}_i(\mathbf{W})/\tau$ delivers the highest critical value amongst the DGPs contained within the weak instruments set or – when the bound $\mathcal{B}_i(\mathbf{W})$ is not sharp – one that is at least as large. As a result, the asymptotic

[10] We adopt the notational convention of Muirhead (1982) for the noncentral Wishart distribution.

rejection rate based on $x = \mathcal{B}_i(\mathbf{W})/\tau$ does not exceed the desired nominal size α for any DGP in the weak instrument set. A similar result applies using $\mathcal{B}_i^s(\hat{\mathbf{W}})$, since the resulting critical values are weakly larger.

While Proposition 3 establishes the size of a testing procedure based on $g_{\min}$, the critical value is infeasible. Although ζ has a noncentral Wishart distribution, critical values for $g_{\min}$ require the distribution of mineval$\{R'_{N,K}(\zeta \otimes I_K)R_{N,K}\}$, which is the minimum eigenvalue of the $N \times N$ matrix consisting of the traces of the $K \times K$ partitions of ζ. To the best of our knowledge, the distribution of such a function of ζ is unknown. Moreover, this distribution depends in general on all parameters in Σ and Ω, not just on $\lambda_{\min}$.

To address both of these challenges, we obtain approximate critical values for the limiting distribution of $g_{\min}$. Specifically, we first derive upper bounds for the second and third cumulants that only depend on $\lambda_{\min}$ and $\mathbf{W}_2$. Next, we consider the class of approximating distributions proposed by Imhof (1961) that match the first three cumulants of an unknown target distribution. We select the Imhof distribution with the largest critical value at significance level α subject to the constraints that the first cumulant matches that of the target distribution and that the second and third cumulants respect the analytical upper bounds on the cumulants of the limiting distribution of $g_{\min}$. The resulting critical value is guaranteed to be conservative relative to that of the Imhof approximation based on the true unknown cumulants of the limiting distribution of $g_{\min}$. The resulting critical values are not provably uniformly conservative for the true limiting distribution, but simulations find the Imhof approximation to be highly accurate.

The first step in our procedure is to derive upper bounds on the cumulants that are free of nuisance parameters. We first discuss the $N = 1$ case, where we can rely in part on existing results in the literature. When $N = 1$, $R'_{N,K}(\zeta \otimes I_K)R_{N,K} = \text{Tr}(\zeta)$ is a scalar. The trace of a noncentral Wishart ζ is a linear combination of noncentral χ^2 variables. While there is no tractable formula for its probability distribution that we are aware of, Mathai (1980) provides an analytical expression for the n-th order cumulant of $\text{Tr}(\zeta)$,

$$\kappa_n = 2^{n-1}(n-1)!\big(\text{Tr}(\Sigma^n) + n\,\text{Tr}(\Sigma^n\Omega)\big) \,. \tag{12}$$

The mean is $\kappa_1 = K(1 + \lambda_{\min})$, since $\text{Tr}(\Sigma) = K$ and $\text{Tr}(\Sigma\Omega) = K\,\text{Tr}(\Lambda) = K\Lambda = K\lambda_{\min}$ when $N = 1$. For $n > 1$, the cumulants are bounded by

$$\kappa_n \leq 2^{n-1}(n-1)!\big(\text{Tr}(\Sigma^n) + nK\lambda_{\min}\,\text{maxeval}\{\Sigma\}^{n-1}\big) \,, \tag{13}$$

which follows from the fact that for $U, V \in \mathbb{P}$, $|\operatorname{Tr}(UV)| \leq \operatorname{maxeval}\{U\} \operatorname{Tr}(V)$, see Fact 8.12.29 in Bernstein (2009), and the fact that $\operatorname{Tr}(\Sigma^n \Omega) \geq 0$. The bounds in (13) only depend on $\lambda_{\min}$ and on Σ, which only depends on $\mathbf{W}_2$ and can therefore be consistently estimated.

The general case with $N \geq 1$ is more involved, as $g_{\min}$ is now asymptotically distributed as the minimum eigenvalue of a matrix containing the traces of the $K \times K$ partitions of ζ. Analogously to Stock and Yogo (2005), we consider the distribution of $\gamma' R'_{N,K}(\zeta \otimes I_K) R_{N,K} \gamma \geq \operatorname{mineval}\{R'_{N,K}(\zeta \otimes I_K) R_{N,K}\}$ as a bounding distribution, where γ is the eigenvector associated with the minimum eigenvalue of $\Lambda = R'_{N,K}(\Sigma\Omega \otimes I_K) R_{N,K}/K$ and $\gamma'\gamma = 1$. The following theorem extends the earlier results to $\gamma' R'_{N,K}(\zeta \otimes I_K) R_{N,K} \gamma$.

Theorem 2. *For $\zeta \sim \mathcal{W}(1, \Sigma, \Omega)$,*

(i) The n-th cumulant of $\gamma' R'_{N,K}(\zeta \otimes I_K) R_{N,K} \gamma$ is

$$\kappa_n = 2^{n-1}(n-1)!\Big(\operatorname{Tr}\left(((\gamma\gamma' \otimes I_K)\Sigma)^n\right) + n \operatorname{Tr}\left(((\gamma\gamma' \otimes I_K)\Sigma)^n \Omega\right)\Big).$$

(ii) The n-th cumulant κ_n with $n > 1$ is bounded by

$$\kappa_n \leq 2^{n-1}(n-1)!\Big(\operatorname{maxeval}\{R'_{N,K}(\Sigma^n \otimes I_K) R_{N,K}\} + nK\lambda_{\min} \operatorname{maxeval}\{\Sigma\}^{n-1}\Big).$$

Proof. See Appendix B. □

Part (i) of the theorem generalizes Mathai (1980) to provide analytical expressions for cumulants of the random scalar $\gamma' R'_{N,K}(\zeta \otimes I_K) R_{N,K} \gamma$. Part (ii) provides upper bounds on the cumulants that only depend on $\lambda_{\min}$ and the consistently estimable Σ, such that they can be used to construct a bounding pivotal distribution. The bounds also nest those for the $N = 1$ case above.

For the mean, $n = 1$, the expression in part (i) of the theorem again simplifies to $\kappa_1 = K(1 + \lambda_{\min})$, which depends only on $\lambda_{\min}$. The mean can, therefore, be targeted exactly, such that only the second and third cumulant of the approximating bounding distribution potentially differ from those of the true distribution. In the special case of the CHSU model, $\Sigma = \mathcal{S} = I_{NK}$, and the bounds in (ii) simplify to

$$2^{n-1}(n-1)!\Big(K + nK\lambda_{\min}\Big). \tag{14}$$

These are the cumulants of a noncentral chi-squared distribution with K degrees of freedom and noncentrality parameter $K\lambda_{\min}$, which is indeed the

exact bounding distribution used by the Stock and Yogo (2005) test.

Next, we define the Imhof (1961) approximating distribution matching the first three cumulants $\kappa_n,\ n = 1, 2, 3$ of the target distribution.

Definition 5. *The Imhof (1961) distribution is*

$$\mathcal{F}_{\mathcal{I}}(x; \kappa_1, \kappa_2, \kappa_3) = \mathcal{F}_{\chi^2_\nu}((x - \kappa_1)4\omega + \nu), \quad \nu = 8\kappa_2\omega^2, \quad \omega = \kappa_2/\kappa_3,$$

where $\mathcal{F}_{\chi^2_\nu}(\cdot)$ is the cdf of the chi-squared distribution with ν degrees of freedom.

Denote the upper bounds on the second and third cumulants in Theorem 2(*ii*) as $\bar{\kappa}_2, \bar{\kappa}_3$. The following proposition is of practical use for our test.

Proposition 4. *There exists α^* such that $\forall\ \alpha \leq \alpha^*$,*

$$\mathcal{F}_{\mathcal{I}}{}^{-1}(1 - \alpha; \kappa_1, \bar{\kappa}_2, \bar{\kappa}_3) \geq \mathcal{F}_{\mathcal{I}}{}^{-1}(1 - \alpha; \kappa_1, \kappa_2', \kappa_3') \quad \forall \kappa_2' \leq \bar{\kappa}_2, \kappa_3' \leq \bar{\kappa}_3.$$

Proof. See Online Appendix **??**. □

The proposition states that, for percentiles far enough in the right tail, the critical values from the Imhof distribution based on κ_1 and the upper bounds $\bar{\kappa}_2, \bar{\kappa}_3$, which are given by $\mathcal{F}_{\mathcal{I}}{}^{-1}(1 - \alpha; \kappa_1, \bar{\kappa}_2, \bar{\kappa}_3)$, are weakly larger, and thus conservative, compared to those from the Imhof distribution based on any other values of the cumulants that satisfy the bounds. In other words, for low enough values of α, the Imhof approximation based on the bounds of Theorem 2(ii) gives the most conservative critical values out of all Imhof distributions satisfying the bounds. In simulations, we found that for a conventional choice of $\alpha = 0.05$, setting the cumulants to their upper bounds virtually always yields the most conservative critical value. Our code, however, always checks whether the Kuhn-Tucker conditions of the associated maximization problem are satisfied at the upper bounds. If this is not the case, which may happen, for example, for $\alpha = 0.10$, the code solves numerically for the most conservative critical value respecting the bounds, which is a relatively straightforward optimization problem.

As discussed in Section 1, the critical values generated by our test under the relative bias criterion are essentially identical to those of Montiel Olea and Pflueger (2013) when $N = 1$ and $K > 2$. If one imposes the Kronecker structure on $\mathbf{W}$, the absolute and relative versions of the tests coincide exactly, and the critical values are, in that case, very similar to those tabulated by Stock and Yogo (2005). The numerical differences with the Stock and Yogo (2005) values are essentially entirely due to the Nagar approximation since the differences between the Imhof approximation with the exact cumulants in

(14) and the noncentral chi-squared bounding distribution are very small, see also Solomon and Stephens (1977) or Bodenham and Adams (2015). Online Appendix **??** illustrates the critical values that arise for a range of models and under either bias criteria. The same Appendix also discusses in greater depth how our critical values compare to those from existing tests.

2.5 Modifications for Models with $K \leq N+1$

Plausible instruments are often hard to come by, and models with only N or $N+1$ instruments are, therefore, empirically highly relevant. Stock and Yogo (2005) do not report critical values for $K \leq N+1$ since the bias criterion, B_i, does not exist when $K = N$. Depending on assumptions, the bias also does not exist when $K = N+1$ or it may be difficult to approximate accurately. In contrast, the bias criterion under the Nagar approximation, $B_{i,n}$, always exists without further assumptions even when $K \leq N+1$, and it is trivial to evaluate numerically. This motivates Montiel Olea and Pflueger (2013) to expand the use of their test to models with $N = 1$ and $K \leq 2$ without any modifications. However, if the true bias does not exist, or the Nagar approximation is very poor, it is reasonable to question the usefulness of a test based on the Nagar bias. As we explain next, we modify our generalized test to deal with the problematic $K \leq N+1$ models in a more satisfactory manner.

We first discuss models with a degree of overidentification equal to one, $K = N+1$. In these models, the bias does, in fact, exist provided that u_t and v_t are Gaussian, as discussed in some detail in Kinal (1980) or Skeels and Windmeijer (2018). The latter authors compute the bias analytically for the CHSU model with $N = 1, K = 2$. The additional Gaussianity assumption is a price many researchers are likely willing to pay for a bias-based weak instruments test. Even so, the accuracy of the Nagar approximation can be very poor for $N = K+1$ models. The CHSU model is a particularly dramatic example of how severe the approximation error can be. When $\mathbf{W}$ has the Kronecker form, the worst-case Nagar bias in Theorem 1 simplifies to $B^*_{i,n}(\mathbf{W}, \lambda_{\min}) = \lambda^{-1}_{\min}|K-(N+1)|/K$. For $K = N+1$, this means the worst-case Nagar bias is always exactly zero! As a result, for $N = 1$ the Montiel Olea and Pflueger (2013) test with $\alpha = 0.05$ and $\tau = 0.10$ yields a critical value of 3.00, whereas the true value obtained analytically by Skeels and Windmeijer (2018) is 7.85. Based on the results of simulations from millions of DGPs, which are discussed below and in Online Appendix **??**, accuracy concerns can also arise for general $\mathbf{W}$ in models with $N = K+1$.

Given these problems, our solution is to always use the most conservative of the two terms in the alternative upper bound in part (ii) of Theorem 1 instead of the tighter (and often sharp) upper bound in part (i). While still motivated by the analytical expression for the Nagar bias, the more conservative bound results in larger critical values that, in our simulations across millions of DGPs, prove very effective at controlling the true bias at the nominal level while also retaining meaningful power, see Online Appendix **??**. We therefore recommend using the more conservative bound for models with $K = N + 1$, which is why our critical values for the relative bias are larger than those of Montiel Olea and Pflueger (2013) for $N = 1$ and $K = 2$.

In just-identified models, $K = N$, the 2SLS bias generally does not exist. Rather than relying on the existence of the Nagar bias, we instead follow Andrews and Armstrong (2017) and Angrist and Kolesár (2024) and focus on the median 2SLS bias instead. The Nagar approximation can easily be adapted to analytically approximate the median bias for just-identified models with a single endogenous regressor. When $K = N = 1$, the one non-zero term in the second-order approximation depends on the random scalar ψ^2, where ψ has a standard normal distribution such that $E[\psi^2] = 1$ and median $(\psi^2) = 0.455$. Therefore, simply multiplying the Nagar bias by median$(\chi^2_1)/E[\chi^2_1] = 0.455$ yields a Nagar approximation of the median bias. In practice, a test based on the median bias of 2SLS can then be implemented with our testing procedure simply by re-scaling the tolerance, $\tau_{med} = \tau/0.455$, where τ is the desired tolerance of the test. The accompanying code makes this adjustment automatically for $K = N = 1$ models. The resulting critical value for the relative bias test is always smaller than that of Montiel Olea and Pflueger (2013) for $N = K = 1$, although it is important to keep in mind that our test controls median bias rather than the mean bias in this case.

Unfortunately, for $K = N > 1$, we are unable to obtain a similarly tractable Nagar approximation to the median bias. Our solution is the same as for $K = N + 1$ models: the test uses the most conservative term in the alternative upper bound in part (ii) of Theorem 1. As the simulations from millions of DGPs in Online Appendix **??** show, this is a highly effective bound for the median bias in just-identified models with a range of dimensions, and the test remains powerful.

We acknowledge that, except for $K = N = 1$, our justification for the alternative bounds for $K \leq N+1$ models relies on simulation evidence rather than analytical results. At the same time, unless researchers are truly only concerned with the Nagar bias, the simulation evidence is clear that modifications like these are necessary to effectively control the actual mean or

median bias in those models. Applications with $K \leq N+1$ are common, and first-stage test statistics are routinely reported. For a bias-based weak instruments test, the critical values under the above modifications have a stronger justification than any others currently available for those models.

2.6 Summary of `gweakivtest_critical_values.m`

We conclude the discussion of our test with a description of the Matlab function `gweakivtest_critical_values.m`. To compute the critical values, the required inputs are (a) an estimate of the robust covariance of the reduced-form and first-stage coefficients, $\hat{\mathbf{W}}$, (b) the choice of scaling, either `abs` or `rel`, (c) the significance level for the test α, and (d) bias tolerance level τ, and (e) an estimate of the robust covariance of reduced-form and first-stage errors, $\hat{\Sigma}_{wv}$, in the case `abs` is selected.

1. If $K > N+1$, the code obtains the threshold value $\lambda^*_{\min}(\tau) = \mathcal{B}_i(\hat{\mathbf{W}})/\tau$ where $\mathcal{B}_i(\hat{\mathbf{W}})$ is calculated numerically using the optimization algorithm of Wen and Yin (2013). The algorithm is not guaranteed to find the global optimum. Unless the user specifies otherwise, our code takes the maximum over the optima found for 1000 starting values generated by N columns of $K \times K$ matrices drawn from the Haar distribution, i.e. uniformly sampled from the space of all orthonormal matrices. If the user selects the simplified testing option, the threshold value is set to $\lambda^*_{\min}(\tau) = \mathcal{B}^s_i(\hat{\mathbf{W}})$. If $K \leq N+1$, the conservative bounds described in Section 2.5 are used; if $K = N = 1$, τ is replaced with τ_{med} to provide a more powerful test for the median bias.

2. Next, the code calculates the upper bounds on the second and third cumulants of the approximate limiting distribution of the test statistic $g_{\min}$,

$$\kappa^*_2 = 2\big(\text{maxeval}\{(I_N \otimes \text{vec}(I_K))'((\hat{\Sigma}^2 \otimes I_K)(I_N \otimes \text{vec}(I_K))\} + 2\lambda^*_{\min}(\tau)K\,\text{maxeval}\{\hat{\Sigma}\}\big), \tag{15}$$

$$\kappa^*_3 = 8\big(\text{maxeval}\{(I_N \otimes \text{vec}(I_K))'((\hat{\Sigma}^3 \otimes I_K)(I_N \otimes \text{vec}(I_K))\} + 3\lambda^*_{\min}(\tau)K\,\text{maxeval}\{\hat{\Sigma}\}^2\big), \tag{16}$$

   where $\hat{\Sigma}$ is defined as in Proposition 2 after replacing $\mathbf{W}$ with $\hat{\mathbf{W}}$.

3. The code checks whether the relevant Kuhn-Tucker conditions are satisfied at $\kappa_1 = K(1+\lambda^*_{\min}(\tau))$, $\bar{\kappa}_2 = \kappa^*_2$ and $\bar{\kappa}_3 = \kappa^*_3$, see Online Appendix **??**. If not, the code solves for the values κ'_2 and κ'_3 that do satisfy the constrained optimality conditions using the bounds as starting values.

4. For significance level α, the code computes the critical value from the limiting distribution using the Imhof (1961) distribution in Definition 5. This critical value is divided by K to obtain the critical value that can be compared to the test statistic $g_{\min}$.

3 Extensions to Other Hypothesis Tests

This section presents two extensions of our weak instruments test to other hypotheses of interest in models with multiple endogenous regressors. The first extension is to cases where researchers are concerned with the bias in only a single coefficient. The second extension is a test of near-rank deficiency as another form of failure of instrument relevance.

3.1 Weak Instruments Test for Individual Elements of $\hat{\beta}_{2SLS}$

Our test for weak instruments based on the bias criterion for the full $\hat{\beta}_{2SLS}$ vector can easily be modified to conduct a test based on the bias of a single element of $\hat{\beta}_{2SLS}$. To do that, in the bias criterion in Definition 2, we replace $E[\beta^*_{2SLS}]$ with $e_j^{N\prime}E[\beta^*_{2SLS}]$ and the weighting matrix Ξ_i with $e_j^{N\prime}\Xi_i e_j^N$, where e_j^N is an $N \times 1$ vector with the j-th element equal to one and zeros in all other rows. The resulting bias criterion, B_i^j, effectively puts zero weight on all coefficients except for the one in the j-th position. The weak instrument set, $\mathbb{B}^j_{i\tau}(\mathbf{W})$, is as in Definition 3 after replacing B_i with B_i^j. Note that we maintain Assumption 1 as before, i.e., the first-stage parameters remain local to zero for all endogenous regressors.

Corollary 1. $\sup_{\mathbb{B}^j_{i,\tau}(\mathbf{W})} \lim_{T\to\infty} \text{Prob}\left(g_{\min} > c\left(\alpha, \hat{\mathbf{W}}_2, \mathcal{B}_i(\hat{\mathbf{W}})/\tau_i^j\right)\right) \leq \alpha,$

where

$$\tau^j_{abs} = \tau \times \frac{||\Phi^{-\frac{1}{2}}\Sigma_v^{\frac{1}{2}}||_2}{\sqrt{e_j^{N\prime}\Sigma_v e_j^N}\ ||\Phi^{-\frac{1}{2}}e_j^N||_2}\ ,\ \tau^j_{rel} = \tau.$$

Proof. See Online Appendix **??**. □

Corollary 1 states that the weak instruments test for an individual coefficient can be conducted exactly as the test described in the previous section, with a simple adjustment to τ under the absolute bias criterion. The proof in Online Appendix **??** shows that the bounds on the Nagar bias in Theorem 1 remain the same under the relative bias criterion and differ only by a scalar under the absolute bias criterion such that dividing the tolerance level by that scalar is all that is required. Moreover, the test for an individual element of $\hat{\beta}_{2SLS}$

is always based on a sharp upper bound on the Nagar bias for that element, even under the absolute bias criterion.

3.2 Tests Under Local to Rank Reduction of 1 (LRR1)

While most treatments of weak instruments have focused on the local-to-zero setting, where all instruments are uniformly weak, similar problems arise when $Y'P_ZY$ is of non-zero – but close to reduced – rank. Sanderson and Windmeijer (2016) consider a "local-to-rank-reduction-of-one" (LRR1) asymptotic embedding and show how, for the CHSU model, an F-statistic with Stock and Yogo (2005) critical values can be used to conduct bias-based tests for this alternative violation of instrument relevance. In this section, we describe how our test statistic and critical values can be used in a similar fashion in the presence of heteroskedasticity and autocorrelation.

We first formalize the new asymptotic embedding:

Assumption 3. *The j^{th} column of Π is $\Pi_j = \Pi_{-j}\delta + c/\sqrt{T}$, where $c \in \mathbb{R}^K$, $\delta \in \mathbb{R}^{N-1}$, and the matrix $\Pi_{-j} \in \mathbb{R}^{K\times(N-1)}$ containing the remaining $N-1$ columns of Π is of full column rank.*

In this framework, Π is asymptotically not rank zero but rank $N-1$. When $\delta = 0$, this assumption covers cases where instruments are weak for one endogenous regressor, but the instruments are strong for the remaining regressors. It also nests settings where $Y'P_ZY$ is statistically close to singular even when instruments are strong, due to collinearity in Π.

Analogous to Sanderson and Windmeijer (2016), the first-stage effective F-statistic from a transformed model with a single regressor can be used to test for weak instruments. Specifically, let Y_j denote the j-th regressor and Y_{-j} the remaining regressors in Y. Define the transformed variables

$$y^\perp = M_{\hat{Y}_{-j}} y,\ Y_j^\perp = M_{\hat{Y}_{-j}} Y_j,\ Z^\perp = M_{\hat{Y}_{-j}} \tilde{Z}(\tilde{Z}' M_{\hat{Y}_{-j}} \tilde{Z}/T)^{-\frac{1}{2}}, \tag{17}$$

where $M_{\hat{Y}_{-j}} = I_T - P_Z Y_{-j}(Y_{-j}' P_Z Y_{-j})^{-1} Y_{-j}' P_Z$ and $\tilde{Z}$ contains any $K-N+1$ columns of Z. The transformed variables are the residuals from projecting y, Y_j, and $\tilde{Z}$ on the predicted values $\hat{Y}_{-j} = P_Z Y_{-j}$, and they enter an auxiliary IV regression of $y^\perp$ on $Y_j^\perp$ using the $K-N+1$ instruments in $Z^\perp$.

We show in Online Appendix **??** that the absolute bias of the 2SLS estimate in the auxiliary regression with the transformed variables is equal to the absolute bias in the original regression. As a result, the weak instruments test of Section 2 for the auxiliary regression can be used to test the null hypothesis of weak instruments in the original regression with multiple endogenous

regressors. Intuitively, after projecting out the part of Y_j predicted by $\hat{Y}_{-j}$, the covariance between the resulting residual $Y_j^\perp$ and $Z^\perp$ is once more local-to-zero. The Online Appendix also shows that the bias in a single element of $\hat{\beta}_{2SLS}$, defined as in Section 3.1, is equal to the bias in the auxiliary regression up to a consistently estimable constant. The weak instruments test in the auxiliary regression can, therefore, also be used to test for bias in a single coefficient after an adjustment to the tolerance level. Corollary 2 states these results formally.

Corollary 2. *Under Assumptions 2 and 3, where* $\hat{\mathbf{W}}^\perp$ *and* $g_{\min}^\perp$ *are computed for the auxiliary regression of* $y^\perp$ *on* $Y_j^\perp$ *using instruments* $Z^\perp$*,*

$$(i) \quad \sup_{\mathbb{B}_{abs,\tau}(\mathbf{W})} \lim_{T\to\infty} \text{Prob}\left(g_{\min}^\perp > c\left(\alpha, \hat{\mathbf{W}}_2^\perp, \mathcal{B}_i(\hat{\mathbf{W}}^\perp)/\tau\right)\right) \leq \alpha,$$

$$(ii) \quad \sup_{\mathbb{B}^j_{abs,\tau}(\mathbf{W})} \lim_{T\to\infty} \text{Prob}\left(g_{\min}^\perp > c\left(\alpha, \hat{\mathbf{W}}_2^\perp, \mathcal{B}_i(\hat{\mathbf{W}}^\perp)/\tau^{j*}\right)\right) \leq \alpha,$$

where $\tau^{j*} = \tau \times \sqrt{\tilde{\delta}'\hat{\Sigma}_v^\perp\tilde{\delta}}/\sqrt{e_j^{N\prime}\hat{\Sigma}_v e_j^N}$*,* $\tilde{\delta}$ *is such that* $\tilde{\delta}_j = 1$ *and* $\tilde{\delta}_{-j} = -\hat{\delta} = (Y_{-j}'P_Z Y_{-j})^{-1}Y_{-j}'P_Z Y_j$*, and* $\hat{\Sigma}_v^\perp$ *and* $\hat{\Phi}^\perp$ *are consistent estimators based on the auxiliary regression.*

Proof. See Online Appendix **??**. □

Critical values for a test for absolute bias in $\hat{\beta}_{2SLS}$ can be constructed using the testing procedure in Section 2 applied to the auxiliary regression with a single endogenous regressor. The test for absolute bias in $\hat{\beta}_{2SLS,j}$ proceeds similarly, but after multiplying the tolerance level by $\sqrt{\tilde{\delta}'\Sigma_v\tilde{\delta}}/\sqrt{e_j^{N\prime}\Sigma_v e_j^N}$. In both cases, the test is based on the sharp upper bound on the Nagar bias, and $g_{\min}$ for the auxiliary IV regression is the relevant test statistic. The latter is also the effective F-statistic since the auxiliary regression has a single endogenous regressor.

Part (i) in the Corollary effectively extends the test in Sanderson and Windmeijer (2016) to models with general $\mathbf{W}$. As in their test, the results depend on the choice of j in finite samples, although in our simulations the differences are typically very small in practice (e.g., the choice of j in Figure 2b matters little). Part (ii) additionally provides a test for a single element in $\hat{\beta}_{2SLS}$. While Sanderson and Windmeijer (2016) are unable to relate the single-element bias analytically to $\lambda_{\min}$, we are able to do so because of our analytical formula for the bias under the Nagar approximation. Corollary 2 only covers the absolute bias criterion since the ratio of B_{rel} to the relative bias of the auxiliary regression still depends on the unknown parameters in β.

While we focus on rank reductions of one as in Sanderson and Windmeijer (2016), Online Appendix **??** establishes a more general version of Corollary 2 under arbitrary rank reductions between 1 and $N-1$. To the best of our knowledge, such a generalization has not previously been developed in the literature. If, based on their knowledge of the empirical problem and the available instruments, researchers have specific concerns about rank reductions of higher order, they can apply the more general version of Corollary 2 in the Online Appendix. We caution, however, that each version of the test is only valid under the specific assumed asymptotic embedding, and researchers should be careful in selecting the embedding that best captures their specific concerns. Developing a testing procedure that is uniformly valid under all possible rank reductions poses significant additional challenges, and is outside the scope of this paper.

4 Simulations

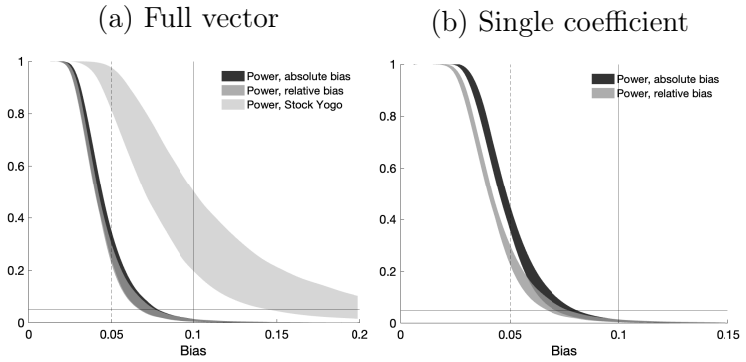
In this section, we evaluate our test in finite-sample simulations using extensions of the design in Sanderson and Windmeijer (2016). We also summarize results from extensive additional simulations based on both asymptotic and empirically calibrated designs, see Online Appendices **??** and **??**.

4.1 Main Setting

Our main simulation designs follow Sanderson and Windmeijer (2016) and accordingly have dimensions $N = 2$, $K = 4$, and $T = 10,000$. The structural and first-stage errors are $[u_t \quad v_t']' \sim \mathcal{N}(0, \Sigma_{uv})$, with $\Sigma_{uv} = \begin{bmatrix} 1 & 0.1 & -0.7 \\ 0.1 & 1 & -0.7 \\ -0.7 & -0.7 & 1 \end{bmatrix}$. Different from Sanderson and Windmeijer (2016), we generate conditional heteroskedasticity by letting $Z_t = (\bar{Z}'\bar{Z}/T)^{-\frac{1}{2}}\bar{Z}_t$ where $\bar{Z}_t \sim \mathcal{N}(0, Q_t)$ and $Q_t = I_K + \Gamma \begin{bmatrix} w_t \\ v_t \end{bmatrix} \begin{bmatrix} w_t & v_t' \end{bmatrix}' \Gamma'$, where Γ is a $K \times (N+1)$ random matrix with elements drawn uniformly on the interval from -1 to 1. Our baseline simulations are based on 10 different realizations of Γ.

We generate data using $\beta = [0.5, -0.3]'$ under both the local-to-zero and LRR1 asymptotic embeddings. For LRR1, we set $\Pi = [\Pi_2\delta + (0, \ a_1)'/\sqrt{T} \ : \ \Pi_2]$ with $\Pi_2 = [-0.5, 0.5, -0.5, 0.5]'$ and $\delta = 0.7$ as in Sanderson and Windmeijer (2016). We vary a_1 to control the value of the concentration parameter and construct power curves. For local-to-zero, we set $\Pi = a_2 \times [\Pi_2\delta + [0, \ 3.1925]' \ : \ \Pi_2]/\sqrt{T}$ using the same values for Π_2 and δ. We control identification strength

Figure 1: Power of Tests in Local-to-Zero Framework



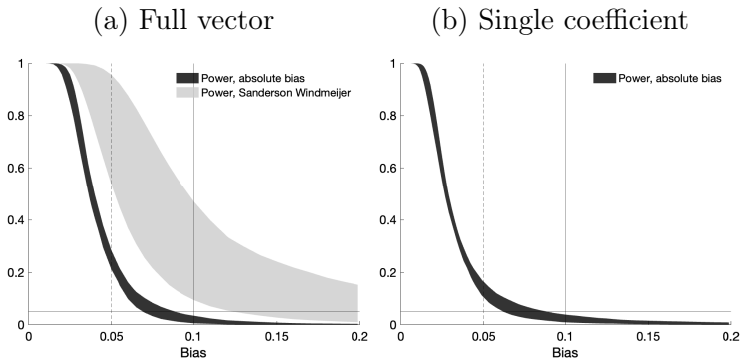
Notes: Power envelopes for various tests with $\tau = 0.10$ and $\alpha = 0.05$ over 10 different local-to-zero simulation designs based on Sanderson and Windmeijer (2016), as described in the text. Areas are the envelopes of power curves for the absolute (blue), relative (red), and Stock and Yogo (2005) (yellow) bias tests. Panel (a) presents tests for the full vector $\hat{\beta}_{2SLS}$ and (b) on a single coefficient $\hat{\beta}_{2SLS,2}$, as described in Sections 2 and 3.1, respectively.

by varying a_2 to match a given minimum eigenvalue of the concentration parameter. We generate 10,000 Monte Carlo samples for each design. For each sample, $\hat{\mathbf{W}}$ is the White (1980) heteroskedasticity-robust variance estimate.

4.2 Local-to-Zero Results

Figure 1 plots the envelopes of power curves for $\tau = 0.10$ and $\alpha = 0.05$ as a function of the worst-case Nagar bias across 10 different values of Γ. Panel (a) presents results for three tests for the bias of the full vector of estimates $\hat{\beta}_{2SLS}$. The yellow area is the Stock and Yogo (2005) test, which incorrectly assumes homoskedasticity. As a result, the test shows large size distortions. When the true worst-case Nagar bias is at $\tau = 0.10$ (solid vertical line), the rejection rates range from 0.197 to 0.503, far exceeding the nominal size of $\alpha = 0.05$ (horizontal line). The blue and red areas in Figure 1 are for the robust tests based on the absolute and relative bias criteria, respectively, as described in Section 2. In contrast to the Stock and Yogo (2005) test, both versions of the robust test control size for all DGPs. Under the absolute bias criterion, for example, the empirical size ranges from 0.004 to 0.015. The use of an approximating distribution that is generally generally bounds the true distribution means that the robust tests are conservative: the null of weak instruments is rejected less frequently than $\alpha = 0.05$ when the true worst-case Nagar bias is 0.10. Nevertheless, both robust tests have nontrivial power. The rejection rates under the absolute criterion, for example, rise to between 0.240 and 0.350 when the true worst-case Nagar bias is 0.05 (dashed vertical line). The relative bias power envelope is very similar to that for absolute bias.

Figure 2: Power of Tests in LRR1 Framework



Notes: Power envelopes for various tests over 10 different LRR1 simulation designs based on Sanderson and Windmeijer (2016), as described in the text. Panel (a) plots tests for bias in the full vector $\hat{\beta}_{2SLS}$, based on the auxiliary regression for $Y_2^{\perp}$, and Panel (b) for bias in $\hat{\beta}_{2SLS,2}$ alone, based on the same regression. Areas are the envelopes of power curves for the absolute bias test (blue) as in Section 3.2, and for the Sanderson and Windmeijer (2016) (yellow) test.

Panel (b) in Figure 1 considers the robust tests for a single coefficient. The power envelope for the relative bias test is unchanged from Panel (a) since $\tau_{rel}^{j} = \tau$, while that of the absolute bias test reflects the required adjusted tolerance level. Size is controlled, but conservative, for both versions of the test and ranges, for example, from 0.004 to 0.014 for the absolute bias test.

4.3 LRR1 Results

Panels (a) and (b) in Figure 2 plot the power envelopes for the LRR1 tests for either the full vector $\hat{\beta}_{2SLS}$, or $\hat{\beta}_{2SLS,2}$, respectively, based on the auxiliary regression for $Y_2^{\perp}$. The results using $Y_1^{\perp}$ (and for $\hat{\beta}_{2SLS,1}$) are similar and are omitted for brevity. The curves for the Sanderson and Windmeijer (2016) tests (yellow) reveal large size distortions when the true worst-case Nagar bias is 0.10 (full vertical line). Because of the incorrect assumption of homoskedasticity, the empirical size is as high as 0.471 in the simulations. In contrast, the robust tests described in Section 3.2 (blue) control size well for both targets, with empirical size ranging between 0.006 and 0.035 for $\hat{\beta}_{2SLS}$ and 0.009 and 0.039 for $\hat{\beta}_{2SLS,2}$. The robust tests have non-trivial power, rising sharply for worst-case Nagar bias below 0.10. At a bias of 0.05 (dashed line), power ranges between 0.211 and 0.284 for $\hat{\beta}_{2SLS}$ and 0.105 and 0.165 for $\hat{\beta}_{2SLS,2}$.

4.4 Additional Simulation Results

The finite-sample simulations based on the Sanderson and Windmeijer (2016) design, extended to feature heteroskedasticity, illustrate the good performance

of our proposed tests in a few particular DGPs. However, given the small range of heteroskedastic structures and eigenstructures of Λ employed in the simulations, the results are far from general. To balance a desire for uniformity with computational feasibility, we report results for an extensive number of asymptotic simulations in Online Appendix **??**. In addition, Online Appendix **??** presents further finite-sample simulation results using smaller samples, clustered error designs, and an empirically-calibrated DGP.

For the asymptotic simulations, we study models with a range of different dimensions, and for each, we consider 5,000,000 randomly drawn DGPs. In half of these DGPs, the parameters are drawn essentially uniformly, and in the other half, they are drawn in the neighborhood of the parameters yielding the worst-case Nagar bias given a random draw of the matrix $\mathbf{W}$. For each DGP, we draw values for $T^{-\frac{1}{2}}Z'w$ and $T^{-\frac{1}{2}}Z'v$ from their asymptotic distributions, taking $\mathbf{W}$ as known, and construct test statistics accordingly.

The main conclusion from the asymptotic simulations is that our testing procedures perform exactly as intended across a large number of DGPs and model dimensions. First, we confirm that, for models with $K > N + 1$, the worst-case Nagar bias is a highly effective upper bound for (Monte Carlo approximations of) the 2SLS bias at the true model parameters. For models with $K = N + 1$, the most conservative bound in part (ii) of Theorem 1 is also a highly effective upper bound for the 2SLS bias, whereas the worst-case Nagar bias in part (i) of Theorem 1 in general is not. The simulations, therefore, support our recommendation for using the conservative bound in those models and for making the same modification when using the Montiel Olea and Pflueger (2013) test when $N = 1$ and $K = 2$. As explained in Section 2.5, in models with $K = N$, we modify the bias criterion to assess the median bias rather than the mean bias. Our simulations demonstrate that the more conservative bound is also a highly effective bound for Monte Carlo median bias in just-identified models.

Second, we find that both versions of the robust test (based on absolute or relative bias) effectively control size against the upper bounds on the Nagar bias. This is the case across all model dimensions we consider, including when $K \leq N + 1$ as long as the modifications discussed in Section 2.5 are adopted.

Third, the robust tests have considerable power in general, with rejection rates that rise sharply not too far below the chosen tolerance level and that go to unity as the bounds on the Nagar bias in the DGP become smaller.

Finally, we explore the performance of t-tests for elements of β conditional on the value of the first-stage test. While developing a general size-based test is beyond the scope of this paper, the simulations demonstrate a strong rela-

tionship between the test statistic $g_{\min}$ and the size distortions of a standard two-side t-statistic for $\hat{\beta}_{2SLS}$. Provided the number of instruments K remains relatively small, size distortions appear well controlled at values of $g_{\min}$ well below those required to control bias at $\tau = 0.10$.

In order to also evaluate the performance of the tests in a small-sample setting that reflects an actual empirical application, Online Appendix **??** presents the simulation results based on a DGP that is calibrated to the empirical application in the next section. The results again confirm that the robust tests effectively control size and attain meaningful power as the bias decreases below the tolerance level.

5 Empirical Application

We illustrate our test in an application due to Ramey and Zubairy (2018), who use local projections to estimate government spending multipliers using military spending news and recursively identified government spending shocks as instruments. The authors' key innovation is to allow the spending multipliers to depend on the state of the business cycle or, alternatively, on the monetary policy regime. To this end, the authors interact government spending with an indicator of whether the economy is in a period of slack or an indicator of whether the policy rate is constrained at the zero lower bound (ZLB). Using their original notation, Ramey and Zubairy (2018) estimate cumulative multipliers for $h = 0, 1, \ldots$ based on

$$\begin{aligned}(18)\quad \sum_{j=0}^{h} y_{t+j} =& I_{t-1}\left[\gamma_{A,h} + \phi_{A,h}(L)z_{t-1} + m_{A,h}\sum_{j=0}^{h} g_{t+j}\right] \\ &+ (1 - I_{t-1})\left[\gamma_{B,h} + \phi_{B,h}(L)z_{t-1} + m_{B,h}\sum_{j=0}^{h} g_{t+j}\right] + \omega_{t+h},\end{aligned}$$

where h is the horizon in quarters, y_t is detrended GDP, I_{t-1} is the regime indicator, z_{t-1} is a vector of controls, $\phi_{A,h}(L)$ and $\phi_{B,h}(L)$ are polynomials in the lag operator, g_t is government spending divided by trend GDP, and $m_{A,h}, m_{B,h}$ are the cumulative spending multipliers over h quarters in the respective states. The use of interaction terms involving endogenous regressors is quite common and is one example of how multiple endogenous regressors often arise in practical applications.

To assess instrument relevance, Ramey and Zubairy (2018) apply the Montiel Olea and Pflueger (2013) test to the individual subsamples implied by the regime indicators, as for each such subsample there is only a single endogenous

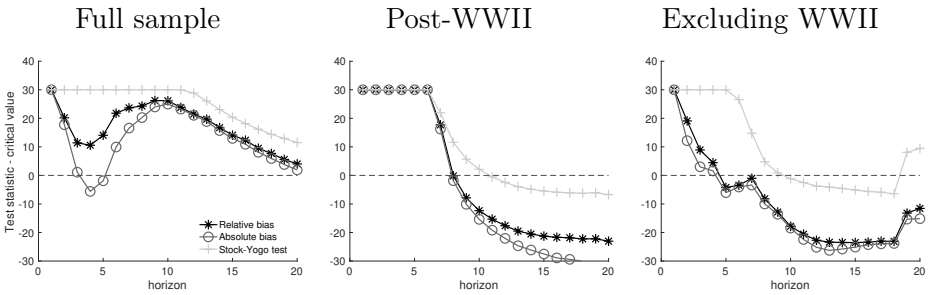
regressor. However, to assess whether the multiplier estimates are statistically different across regimes, government spending ultimately has to be interacted with the indicators in a single specification as in (18), in which there are not one but two endogenous regressors (N=2). As autocorrelated errors are a common feature of local projections such as (18), Ramey and Zubairy (2018) unfortunately cannot rely on the Stock and Yogo (2005) test to assess instrument strength for their regression of interest. Our robust tests, in contrast, allow for a direct test of instrument relevance for the specifications in (18). We therefore implement our test as described in Section 1 for each bias criterion using $\tau = 0.10$ and $\alpha = 0.05$, and using the same Newey and West (1994) estimator as Ramey and Zubairy (2018). The instruments are the military news measure and the recursive spending shocks interacted with the regime indicator as instruments, such that $K = 4$. For reference, we replicate Ramey and Zubairy's (2018) results for the regime subsamples and Montiel Olea and Pflueger's (2013) effective F-statistic in Online Appendix **??**.

Figure 3 reports the results for the main specifications and sample periods discussed in Ramey and Zubairy (2018). Panel (a) covers specifications that interact government spending with a measure of slack based on the unemployment rate. The starred blue line plots the difference between our test statistic $g_{\min}$ and the relative bias critical value for $\tau = 0.10$, $\alpha = 0.05$, truncated at 30 for readability as in Ramey and Zubairy (2018). The circled red line does the same for the absolute bias critical values. Although not reported by Ramey and Zubairy (2018), for illustrative purposes, the crossed yellow line plots the difference between the Cragg and Donald (1993) test statistic and the Stock and Yogo (2005) critical value. Panel (b) in Figure 3 shows the corresponding results for specifications where government spending is interacted with an indicator for whether monetary policy is constrained by the ZLB. For the interested reader, we report the test statistics and critical values for our tests separately in Online Appendix **??**.

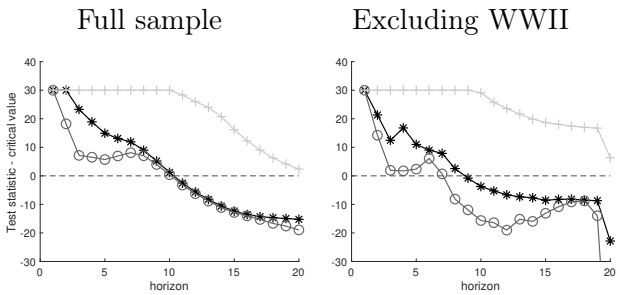
Similar to the regime-specific results reported by Ramey and Zubairy (2018), our robust test with relative bias critical values rejects weak instruments at short horizons across all samples. Bias due to weak instruments becomes a concern in specifications with the slack indicator after between 5 to 8 quarters unless the sample includes WWII. In the specifications with the ZLB indicator, bias becomes a concern for horizons beyond 9 quarters. For the absolute bias version of the test, the conclusions are broadly similar. The key differences are sharp drops in instrument relevance at short horizons (2-4 quarters) for the full sample with the slack indicator and both samples with the ZLB indicator. However, at longer horizons, the results are more similar

Figure 3: Test Results for the Ramey and Zubairy (2018) Regression Across Horizons

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



(b) *Government Spending Interacted with ZLB Indicator*



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 starred blue line plots the difference between the robust test statistic and relative bias critical values for $\tau = 0.10$ and $\alpha = 0.05$ across horizons. As in Ramey and Zubairy (2018), we cap the results at 30 for visibility. The circled red line shows the difference between the robust test statistic and absolute bias critical values. The crossed yellow line shows the difference between the Cragg and Donald (1993) statistic and critical values from Stock and Yogo (2005). 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.

to those based on the relative bias criterion.

Comparison with the results from the Stock and Yogo (2005) test demonstrates the importance of allowing for heteroskedasticity and autocorrelation in the first-stage testing procedures. For three of the five specifications considered in Figure 3, the Stock and Yogo (2005) test leads to a rejection of weak instruments at all horizons considered. For the remaining two specifications, the Stock and Yogo (2005) test rejects weak instruments for an additional 3 to 4 quarters compared to the robust test.

The comparison of our robust test results to the regime-specific results in Ramey and Zubairy (2018), reported in Online Appendix **??**, is also informative. Since the point estimates (and thus the bias) in the interacted regression are the same as those obtained from the regressions in the regime subsamples, it is not surprising that when each regime-specific regression appears to be separately strongly identified, the interacted regression generally is too. This

is particularly true for the relative bias since this test is a direct generalization of the Montiel Olea and Pflueger (2013) test they adopt.[11] The more interesting cases occur when one state is strongly identified, but the other is weakly identified. Ex ante, it is not obvious whether the interacted model would fall above or below the relevant critical value. In this application, the more weakly identified state appears to dictate the test result. While intuitive, this need not be the case in general and depends on the covariance structure of the regressors and instruments.

Finally, Online Appendix **??** reports results from the size-based approach of Andrews (2018) for comparison. While not a weak instruments test per se, the two-step procedure in Andrews (2018) determines the maximum value $\hat{\gamma}$ for which a robust confidence set of size $1-\alpha-\hat{\gamma}$ is contained by non-robust (e.g., Wald) set of size $1-\alpha$. If $\hat{\gamma}$ exceeds a pre-specified tolerance, say 0.10, the conclusion is that instruments are weak. The Online Appendix shows that the size distortions follow a similar qualitative pattern to the bias, growing larger at longer horizons and after omitting WWII or the early part of the sample. With a tolerance level for size distortions of 0.10, weak instruments are rejected much more frequently than the bias-based test, which is also in line with the results from the asymptotic simulations discussed earlier. We emphasize, however, that both tests are informative about different properties of the instruments: one focuses on their ability to recover point estimates, the other on whether they lead to reliable inference using standard methods. A researcher may care about either separately or both.

6 Concluding Remarks

First-stage tests like those proposed by Stock and Yogo (2005) or, more recently, Montiel Olea and Pflueger (2013) and Sanderson and Windmeijer (2016), are a widely-used diagnostic tool to assess the bias of 2SLS. When researchers are not comfortable imposing the assumption of CHSU errors for second-stage inference, they should also avoid imposing such assumptions in first-stage testing procedures. In this paper, we provide generalized weak instruments testing procedures that are valid under heteroskedasticity and autocorrelation regardless of the number of endogenous regressors. Our test is based either on the absolute bias criterion of Stock and Yogo (2005) or on a generalization of the bias criterion of Montiel Olea and Pflueger (2013).

[11] In just a few cases where one state is only marginally strongly identified, the interacted specification is weakly identified. We attribute this to the fact that the test is more conservative for $N > 1$, due to the approximation used to construct a limiting distribution of a minimum eigenvalue. See Online Appendix **??** for a detailed discussion.

It accommodates settings where instruments are weak because the matrix of first-stage parameters is statistically close to zero, as in Stock and Yogo (2005), or because it is near a column rank reduction of one, extending Sanderson and Windmeijer (2016). Finally, we develop extensions to test for the bias in a single element of a vector of 2SLS estimates. The computer code accompanying this paper provides empirical researchers with an easy-to-use bias-based first-stage test under assumptions that match those imposed for second-stage inference. Future work could consider the generalization of the size-based test of Stock and Yogo (2005). The first-stage test statistic in this paper could also be the foundation for extending the t-statistic inference approach for $K = 1, N = 1$ models of Lee et al. (2022). Further, our generalization of the Nagar approximation to the 2SLS bias should also permit extensions to the methods in Ganics et al. (2021) to construct confidence intervals for the 2SLS bias. Finally, given the availability of tests for arbitrary rank reductions, it would be very useful to develop a unified testing procedure that can guide the researcher in correctly selecting the appropriate test.

Data Availability Statement: The data and code underlying this research is available on Zenodo at https://doi.org/10.5281/zenodo.15616807.

References

Anderson, T. W. and Herman Rubin (1949). "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations". In: *The Annals of Mathematical Statistics* 20.1, pp. 46 –63.

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

Andrews, Isaiah and Timothy B. Armstrong (2017). "Unbiased instrumental variables estimation under known first-stage sign". In: *Quantitative Economics* 8.2, pp. 479–503.

Andrews, Isaiah, James H. Stock, and Liyang Sun (2019). "Weak Instruments in Instrumental Variables Regression: Theory and Practice". In: *Annual Review of Economics* 11.1, pp. 727–753.

Angrist, Joshua and Michal Kolesár (2024). "One instrument to rule them all: The bias and coverage of just-ID IV". In: *Journal of Econometrics* 240.2.

Basmann, R. L. (1961). "A Note on the Exact Finite Sample Frequency Functions of Generalized Classical Linear Estimators in Two Leading

Over-Identified Cases". In: *Journal of the American Statistical Association* 56.295, pp. 619–636.

Bernstein, Dennis S. (2009). *Matrix Mathematics, Theory, Facts and Formulas*. Princeton: Princeton University Press.

Bodenham, Dean and Niall Adams (2015). "A comparison of efficient approximations for a weighted sum of chi-squared random variables". In: *Statistics and Computing* 26.

Carrasco, Marine and Mohamed Doukali (2021). "Testing overidentifying restrictions with many instruments and heteroscedasticity using regularised jackknife IV". In: *The Econometrics Journal* 25.1, pp. 71–97.

Cragg, John G. and Stephen G. Donald (1993). "Testing Identifiability and Specification in Instrumental Variable Models". In: *Econometric Theory* 9.2, pp. 222–240.

Ganics, Gergely, Atsushi Inoue, and Barbara Rossi (2021). "Confidence Intervals for Bias and Size Distortion in IV and Local Projections-IV Models". In: *Journal of Business & Economic Statistics* 39.1, pp. 307–324.

Horn, Roger A. and Charles R. Johnson (2013). *Matrix Analysis*. Cambridge: Cambridge University Press.

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

Keane, Michael and Timothy Neal (2022). *A Practical Guide to Weak Instruments*. Discussion Papers 2021-05d. School of Economics, The University of New South Wales.

Kinal, Terrence W. (1980). "The Existence of Moments of k-Class Estimators". In: *Econometrica* 48.1, pp. 241–249.

Kleibergen, Frank and Richard Paap (2006). "Generalized reduced rank tests using the singular value decomposition". In: *Journal of Econometrics* 133.1, pp. 97–126.

Kollo, Tõnu and Dietrich von Rosen (1995). "Approximating by the Wishart distribution". In: *Annals of the Institute of Statistical Mathematics* 47, pp. 767–783.

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.

Letac, Gérard and Hélène Massam (2018). "The Laplace transform and the existence of non-central Wishart distributions". In: *Journal of Multivariate Analysis* 163, pp. 96–110.

Magnus, Jan R. and Heinz Neudecker (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Oxford: John Wiley and Sons Ltd.

Mariano, Roberto S. (1972). "The Existence of Moments of the Ordinary Least Squares and Two-Stage Least Squares Estimators". In: *Econometrica* 40.4, pp. 643–652.

Mathai, Arak (1980). "Moments of the trace of a noncentral Wishart matrix". In: *Communications in Statistics - Theory and Methods* 9, pp. 795–801.

Mikusheva, Anna and Liyang Sun (2021). "Inference with Many Weak Instruments". In: *The Review of Economic Studies* 89.5, pp. 2663–2686.

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.

Muirhead, Robb J. (1982). *Aspects of Multivariate Statistical Theory*. Hoboken, New Jersey: John Wiley and Sons Ltd.

Nagar, A. L. (1959). "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations". In: *Econometrica* 27.4, pp. 575–595.

Newey, Whitney K. and Kenneth D. West (1994). "Automatic Lag Selection in Covariance Matrix Estimation". In: *The Review of Economic Studies* 61.4, pp. 631–653.

Phillips, P. C. B. (1980). "The Exact Distribution of Instrumental Variable Estimators in an Equation Containing n + 1 Endogenous Variables". In: *Econometrica* 48.4, pp. 861–878.

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.

Sanderson, Eleanor and Frank Windmeijer (2016). "A weak instrument F-test in linear IV models with multiple endogenous variables". In: *Journal of Econometrics* 190.2, pp. 212–221.

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

Solomon, Herbert and Michael A. Stephens (1977). "Distribution of a Sum of Weighted Chi-Square Variables". In: *Journal of the American Statistical Association* 72.360, pp. 881–885.

Staiger, Douglas and James H. Stock (1997). "Instrumental Variables Regression with Weak Instruments". In: *Econometrica* 65.3, pp. 557–586.

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.

Wen, Zaiwen and Wotao Yin (2013). "A Feasible Method for Optimization with Orthogonality Constraints". In: *Mathematical Programming* 142, pp. 397–434.

White, Halbert (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". In: *Econometrica* 48.4, pp. 817–838.

Windmeijer, Frank (2025). "The Robust F-statistic as a Test for Weak Instruments". In: *Journal of Econometrics* 247.105951.

A Proof of Theorem 1

Define the function $h : \mathbb{R}^{NK} \mapsto \mathbb{R}^{N\times NK^2}$

$$h(\psi) = \left(R'_{N,K}(\mathcal{S}(l+\psi)(l+\psi)'\mathcal{S}' \otimes I_K)R_{N,K}\right)^{-1} \times R'_{N,K}\left(\mathcal{S}(l+\psi)\psi'\mathcal{S}^{-1} \otimes I_K\right). \tag{A.1}$$

The Nagar approximation of $\mathbf{h}$, denoted by $\mathbf{h}_n$, is the expectation of the second-order Taylor expansion of $Kh(\psi)$ evaluated at $\psi = 0$, and is given in vectorized form by

$$\text{vec}(\mathbf{h}_n) = \frac{K}{2}(\mathbf{1}'_{NK} \otimes I_{(NK)^2}) \left(\sum_{j=1}^{KN}\left(e_j^{NK}(e_j^{NK})' \otimes I_{(NK)^2}\right)(\nabla_2 h(0))_j\right), \tag{A.2}$$

where $\mathbf{1}_m$ is the $m \times 1$ vector of ones, e_j^m the $m \times 1$ vector with j-th element equal to one and zeros in all other rows, and $\nabla_2 h(\psi)$ is the $(NK)^3 \times NK$ second matrix derivative of h. Using the matrix differentiation results in, for instance, Magnus and Neudecker (2019), $\nabla_2 h(\psi)$ is given by

$$\begin{aligned}\nabla_2 h(\psi) =& (I_{NK} \otimes A_3(\psi))\nabla A_1(\psi) + (A_1(\psi)' \otimes I_{(NK)^2})\nabla A_3(\psi) \\ &+ (I_{NK} \otimes A_4(\psi))\nabla A_2(\psi) + (A_2(\psi)' \otimes I_{(NK)^2})\nabla A_4(\psi),\end{aligned} \tag{A.3}$$

where

$$\begin{aligned}
A_1(\psi) &= -(C_1(\psi) \otimes C_1(\psi))B_1\left((\mathcal{S} \otimes \mathcal{S}(l+\psi)) + (\mathcal{S}(l+\psi) \otimes \mathcal{S})\right) \\
A_2(\psi) &= B_2\left(((\mathcal{S}^{-1})' \otimes \mathcal{S}(l+\psi)) + ((\mathcal{S}^{-1})'\psi \otimes \mathcal{S})\right) \\
A_3(\psi) &= \left(\mathcal{S}(l+\psi)\psi'\mathcal{S}^{-1} \otimes I_K\right)' R_{N,K} \otimes I_N \\
A_4(\psi) &= I_{NK^2} \otimes C_1(\psi) \\
\nabla A_1(\psi) &= \left(((\mathcal{S} \otimes \mathcal{S}(l+\psi)) + (\mathcal{S}(l+\psi) \otimes \mathcal{S}))' B_1' \otimes I_{N^2}\right) C_2(\psi) - (I_{NK} \otimes (C_1(\psi) \otimes C_1(\psi)))B_3 \\
\nabla A_2(\psi) &= (I_{NK} \otimes B_2)\left((\text{vec}((\mathcal{S}^{-1})') \otimes I_{NK})\mathcal{S} + (\mathcal{K}_{NK,NK} \otimes I_{NK})(I_{NK} \otimes \text{vec}(\mathcal{S}))(\mathcal{S}^{-1})'\right)
\end{aligned}$$

$$
\begin{aligned}
\nabla A_3(\psi) &= (I_N \otimes \mathcal{K}_{N,NK^2} \otimes I_N)(I_{(NK)^2} \otimes \mathrm{vec}(I_N))\mathcal{K}_{N,NK^2} A_2(\psi) \\
\nabla A_4(\psi) &= (I_{NK^2} \otimes \mathcal{K}_{N,NK^2} \otimes I_N)(\mathrm{vec}(I_{NK^2}) \otimes I_{N^2}) A_1(\psi) \\
B_1 &= (R'_{N,K} \otimes R'_{N,K})(I_{NK} \otimes \mathcal{K}_{K,NK} \otimes I_K)\left(I_{(NK)^2} \otimes \mathrm{vec}(I_K)\right) \\
B_2 &= (I_{NK^2} \otimes R'_{N,K})(I_{NK} \otimes \mathcal{K}_{K,NK} \otimes I_K)\left(I_{(NK)^2} \otimes \mathrm{vec}(I_K)\right) \\
B_3 &= (I_{NK} \otimes B_1)\left(\mathrm{vec}(\mathcal{S} \otimes I_{NK})\mathcal{S} + (\mathcal{K}_{NK,NK} \otimes I_{NK})(I_{NK} \otimes \mathrm{vec}(\mathcal{S}))\mathcal{S}\right) \\
C_1(\psi) &= \left(R'_{N,K}(\mathcal{S}(l+\psi)(l+\psi)'\mathcal{S}' \otimes I_K)R_{N,K}\right)^{-1} \\
C_2(\psi) &= (I_N \otimes \mathcal{K}_{N,N} \otimes I_N)\left[\mathrm{vec}(A_0(\psi)) \otimes I_{N^2} \ : \ I_{N^2} \otimes \mathrm{vec}(A_0(\psi))\right]\left[A_1(\psi)' \ : \ A_1(\psi)'\right]'.
\end{aligned}
$$

Writing (A.2) in matrix form and simplifying yields

$$
\mathbf{h}_n = \Lambda^{-1}\Big(R'_{N,K} - (\mathrm{vec}(\Lambda^{-1}) \otimes I_N)'(I_N \otimes \mathcal{K}_{N,N})\big((I_{N^2} + \mathcal{K}_{N,N})(I_N \otimes L) \otimes L\big)\Big), \tag{A.4}
$$

where $L = K^{-\frac{1}{2}} R'_{N,K}(\mathcal{S}l \otimes I_K)$ and Λ is the concentration parameter in Definition 1. To proceed, we reparametrize the functional dependence of the bias on l through

$$
l = \mathcal{S}^{-1}\sqrt{K}\,\mathrm{vec}(L'_0 \mathcal{D}_\Lambda^{\frac{1}{2}} Q'_\Lambda), \tag{A.5}
$$

where $Q_\Lambda \in \mathbb{O}^{N\times N}, \mathcal{D}_\Lambda \in \mathbb{R}^{N\times N}$ contain the eigenstructure of the concentration parameter Λ, and $L_0 \in \mathbb{O}^{N\times K}$ is an orthogonal matrix. By definition, $\Lambda = Q_\Lambda \mathcal{D}_\Lambda Q'_\Lambda$, where $\mathcal{D}_\Lambda$ is a diagonal matrix containing the eigenvalues $\lambda_i > 0$, $i = 1, ..., N$, and $Q_\Lambda Q'_\Lambda = I_N$. The reparametrization in (A.5) reformulates the choice of the NK parameters of l as an equivalent choice of the N free parameters in $\mathcal{D}_\Lambda$, the $N^2 - (N+1)N/2$ free parameters of Q_Λ, and the $NK - (N+1)N/2$ free parameters of L_0.

Using the eigenvalue decomposition $\Lambda = Q_\Lambda \mathcal{D}_\Lambda Q'_\Lambda$, and the fact that (A.5) implies $L_0 = Q'_\Lambda \Lambda^{-\frac{1}{2}} L$, (A.4) can be rewritten as

$$
\mathbf{h}_n = Q_\Lambda \mathcal{D}_\Lambda^{-\frac{1}{2}} M_1 (\mathcal{D}_\Lambda^{-\frac{1}{2}} Q_\Lambda \otimes L_0 \otimes L_0) M_2, \tag{A.6}
$$

where $M_1 = R'_{N,N}\left(I_{N^3} + (\mathcal{K}_{N,N} \otimes I_N)\right)$ and $M_2 = R_{N,K} R'_{N,K}/(N+1) - I_{NK^2}$, which provides the Nagar approximation for the expectation in the general bias, $\mathbf{h}_n$, with

$$
B_{i,n}(\beta, Q_\Lambda, \mathcal{D}_\Lambda, L_0, \mathbf{W}) = ||\tilde{\Xi}_i^{\frac{1}{2}} \mathbf{h}_n \rho_i||_2 \leq ||\tilde{\Xi}_i^{\frac{1}{2}}||_2 ||\mathbf{h}_n \rho_i||_2. \tag{A.7}
$$

Let $B^*_{i,n}(\mathbf{W}, \lambda_{\min}) = ||\tilde{\Xi}_i^{\frac{1}{2}}||_2 \sup_{\beta, Q_\Lambda, \mathcal{D}_\Lambda, L_0} \{||\mathbf{h}_n \rho_i||_2\}$ denote the upper bound on the Nagar bias over $\beta \in \mathbb{R}^N$, $Q_\Lambda \in \mathbb{O}^{N\times N}$, $L_0 \in \mathbb{O}^{N\times K}$ and $\mathcal{D}_\Lambda$ in the set of all diagonal matrices with no diagonal element smaller than $\lambda_{\min}$, the smallest eigenvalue of the concentration parameter.

Using the definitions of $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_{12}$ in (8),

$$\begin{aligned}(A.8)\quad \text{Tr}(\mathbf{S}_1) &= \text{Tr}((\tilde{\beta}' \otimes I_K)\mathbf{W}(\tilde{\beta} \otimes I_K)) = \tilde{\beta}' R'_{N+1,K}(\mathbf{W} \otimes I_K)R_{N+1,K}\tilde{\beta}, \\ \text{vec}(\mathbf{S}_{12})' &= \text{vec}((\tilde{\beta}' \otimes I_K)[\mathbf{W}_{12} : \mathbf{W}_2])' \\ &= \text{vec}(\tilde{\beta}' R'_{N+1,K}([\mathbf{W}_{12} : \mathbf{W}_2] \otimes I_K))',\end{aligned}$$

and $\sigma_u^2 = \tilde{\beta}'\Sigma_{wv}\tilde{\beta}$, where $\tilde{\beta} = [1 : -\beta']'$. Substituting into the definition of ρ_i in Lemma 1 yields

$$(A.9)\qquad \rho_i = K^{-\frac{1}{2}}\Psi_i x_i / \sqrt{x_i' x_i},$$

where $\Psi_{abs} = \left(\mathcal{S}\mathbf{W}_2^{-\frac{1}{2}}[\mathbf{W}_{12} : \mathbf{W}_2]' \otimes I_K\right)R_{N+1,K}\Sigma_{wv}^{-\frac{1}{2}}$, $x_{abs} = \Sigma_{wv}^{\frac{1}{2}}\tilde{\beta}$, $\Psi_{rel} = \left(\mathcal{S}\mathbf{W}_2^{-\frac{1}{2}}[\mathbf{W}_{12} : \mathbf{W}_2]' \otimes I_K\right)R_{N+1,K}(R'_{N+1,K}(\mathbf{W} \otimes I_K)R_{N+1,K})^{-\frac{1}{2}}$, and $x_{rel} = (R'_{N+1,K}(\mathbf{W} \otimes I_K)R_{N+1,K})^{\frac{1}{2}}\tilde{\beta}$. Since

$$(A.10)\qquad \sup_{\beta \in \mathbb{R}^N} ||\mathbf{h}_n \rho_i||_2 = K^{-\frac{1}{2}} \sup_{x_i \in \mathbb{R}^{N+1}} \frac{||\mathbf{h}_n \Psi_i x_i||_2}{||x_i||_2} = K^{-\frac{1}{2}}||\mathbf{h}_n \Psi_i||_2,$$

the optimization of the Nagar bias over β amounts to taking the largest singular value of the matrix $\mathbf{h}_n\Psi_i$.

Next, note that $\mathbf{h}_n\mathbf{h}_n' = Q_\Lambda \mathcal{D}_h Q_\Lambda'$ where $\mathcal{D}_h = (K - 2(1+N))\mathcal{D}_\Lambda^{-2} + \mathcal{D}_\Lambda^{-\frac{1}{2}} M_1(\mathcal{D}_\Lambda^{-1} \otimes I_{N^2})M_1'\mathcal{D}_\Lambda^{-\frac{1}{2}}$ is a diagonal matrix, such that Q_Λ are eigenvectors of $\mathbf{h}_n\mathbf{h}_n'$ and $\mathcal{D}_h$ contains the eigenvalues. The j-th diagonal element of $\mathcal{D}_h$ is

$$(A.11)\qquad \frac{1}{\lambda_j^2}\left(K - N + 1 + \sum_{k \neq j}^{N} \frac{\lambda_j}{\lambda_k}\right) > 0,$$

where $\lambda_j > 0$ is the j-th eigenvalue of Λ. Each eigenvalue of $\mathbf{h}_n\mathbf{h}_n'$ is decreasing in all eigenvalues of Λ. Making the dependence on $\mathcal{D}_\Lambda$ explicit by the notation $\mathbf{h}_n(\mathcal{D}_\Lambda)$ and fixing Q_Λ, it is therefore the case that $\lambda_{\min}^{-2}\mathbf{h}_n(I_N)'\mathbf{h}_n(I_N) - \mathbf{h}_n(\mathcal{D}_\Lambda)'\mathbf{h}_n(\mathcal{D}_\Lambda)$ is positive semidefinite for all $\mathcal{D}_\Lambda$ with $\lambda_{\min}$ as the smallest diagonal element. This, in turn, implies that $\lambda_{\min}^{-2}\Psi_i'\mathbf{h}_n(I_N)'\mathbf{h}_n(I_N)\Psi_i - \Psi_i'\mathbf{h}_n(\mathcal{D}_\Lambda)'\mathbf{h}_n(\mathcal{D}_\Lambda)\Psi_i$ is positive semidefinite, see Proposition 8.1.2 in Bernstein (2009). It follows from Weyl's inequality that $\lambda_{\min}^{-1}||\mathbf{h}_n(I_N)\Psi_i||_2 \geq ||\mathbf{h}_n(\mathcal{D}_\Lambda)\Psi_i||_2$, see for example Theorem 8.4.9 in Bernstein (2009). Therefore,

$$(A.12)\qquad \sup_{\mathcal{D}_\Lambda : \lambda_j \geq \lambda_{\min}} K^{-\frac{1}{2}}||\mathbf{h}_n(\mathcal{D}_\Lambda)\Psi_i||_2 = K^{-\frac{1}{2}}\lambda_{\min}^{-1}||Q_\Lambda M_1(Q_\Lambda \otimes L_0 \otimes L_0)M_2\Psi_i||_2,$$

which states that the largest bias occurs when all eigenvalues of the concentration parameter are equal to the smallest eigenvalue, and therefore when $\mathcal{D}_\Lambda = \lambda_{\min} I_N$.

Finally, $\sup_{L_0 \in \mathbb{O}^{N\times K}} \{||Q_\Lambda M_1(Q_\Lambda \otimes L_0 \otimes L_0)M_2\Psi_i||_2\} = \sup_{L_0 \in \mathbb{O}^{N\times K}} \{||M_1(I_N \otimes L_0 \otimes L_0)M_2\Psi_i||_2\}$ for any Q_Λ, and therefore the upper bound for the bias does not depend on Q_Λ. This means the upper bound is

$$B^*_{i,n}(\mathbf{W}, \lambda_{\min}) = \lambda_{\min}^{-1}||\tilde{\Xi}_i^{\frac{1}{2}}||_2 K^{-\frac{1}{2}} \sup_{L_0 \in \mathbb{O}^{N\times K}} \{||M_1(I_N \otimes L_0 \otimes L_0)M_2\Psi_i||_2\}, \tag{A.13}$$

which concludes the proof of part (i) of the theorem.

Turning to part (ii), the upper bound $B^*_{i,n}(\mathbf{W}, \lambda_{\min}) \leq \lambda_{\min}^{-1}||\tilde{\Xi}_i^{\frac{1}{2}}||_2(2(N+1)/K)^{\frac{1}{2}}||M_2\Psi_i||_2$ follows from $K^{-\frac{1}{2}}||\mathbf{h}_n(\lambda_{\min}I_N)\Psi_i||_2 \leq K^{-\frac{1}{2}}\lambda_{\min}^{-1}||M_1||_2||(I_N \otimes L_0 \otimes L_0)||_2||M_2\Psi_i||_2$ and the fact that $||M_1||_2 = (2(N+1))^{\frac{1}{2}}$ and $||(I_N \otimes L_0 \otimes L_0)||_2 = 1$. The inequality follows from Proposition 9.6.1 in Bernstein (2009). Finally, the upper bound, $B^*_{i,n}(\mathbf{W}, \lambda_{\min}) \leq \lambda_{\min}^{-1}||\tilde{\Xi}_i^{\frac{1}{2}}||_2||\Psi_i||_2$, follows from $K^{-\frac{1}{2}}||\mathbf{h}_n(\lambda_{\min}I_N)\Psi_i||_2 \leq K^{-\frac{1}{2}}||\mathbf{h}_n(\lambda_{\min}I_N)||_2||\Psi_i||_2 = \lambda_{\min}^{-1}||\Psi_i||_2$ since $||\mathbf{h}_n(\lambda_{\min}I_N)||_2 = K^{\frac{1}{2}}\lambda_{\min}^{-1}$, see (A.11).

B Proof of Theorem 2.

The Laplace transform of the trace of a noncentral Wishart distribution is given in Letac and Massam (2018), equation (1), from which it follows that the cumulant generating function is

$$K_{\mathrm{Tr}(\zeta)}(t) = -\frac{1}{2}\mathrm{Tr}\,(\Omega) - \frac{K}{2}\ln|I_{NK} - 2\Sigma| + \frac{1}{2}\mathrm{Tr}\,((I_{NK} - 2\Sigma)^{-1}\Omega) \tag{B.14}$$

for a scalar, t. We follow e.g., Muirhead (1982) and Kollo and Rosen (1995) in evaluating the cumulant generating function for a submatrix,

$$K_{\mathrm{Tr}(\zeta)}(T_N) = -\frac{1}{2}\mathrm{Tr}\,(\Omega) - \frac{K}{2}\ln|I_{NK} - 2M(T_N)\Sigma| + \frac{1}{2}\mathrm{Tr}\,((I_{NK} - 2M(T_N)\Sigma)^{-1}\Omega), \tag{B.15}$$

where T_N is a $N \times N$ matrix and

$$M(T_N) = \sum_{i,j=1,\ldots,N} t_{ij}M_{ij}, \quad M_{ij} = e_j e_i', \tag{B.16}$$

where e_i is the i-th block of K columns of the matrix I_{NK}, so that $M_{ij}\zeta$ is the matrix containing the i-th block of K rows of ζ in its j-th block of K rows, and zero otherwise. Indexing each selection matrix M_{ij} to a scalar value t_{ij} yields the cumulant generating function of the trace of $M_{ij}\zeta$, analogously to Mathai (1980), setting the remainder of T_N to zero. $\mathrm{Tr}(M_{ij}\zeta) = \mathrm{Tr}(\zeta_{ij})$, since the j-th diagonal block of $M_{ij}W$ is ζ_{ij}, and other diagonal blocks are zero.

The ij entry of $R'_{N,K}(\zeta \otimes I_K)R_{N,K}$ corresponds to the trace of the ij $K \times K$ block of ζ. Thus, the n-th cumulants of $R'_{N,K}(\zeta \otimes I_K)R_{N,K}$ are obtained by taking the coefficients on $\frac{T_N^n}{n!}$ in the Taylor expansion of $K_{\mathrm{Tr}(\zeta)}(T_N)$, evaluated at $T_N = 0$. Let $\iota(i)$ denote the index of some $K \times K$ block of a $NK \times NK$ matrix. Then $\kappa_n(\zeta)^{\iota(1),\iota(2),\ldots}$ denotes the n-th cumulant of $\mathrm{Tr}(\zeta_{\iota(1)})$ with $\mathrm{Tr}(\zeta_{\iota(2)})\ldots$ (i.e. the

covariance for $n = 2$, etc.). Taking such partial derivatives yields the following expressions:

$$\kappa_1^{\iota(1)}(\zeta) = \frac{K}{2}\operatorname{Tr}(2M_{\iota(1)}\Sigma) + \frac{1}{2}\operatorname{Tr}(2M_{\iota(1)}\Sigma\Omega) \tag{B.17}$$

$$\kappa_2^{\iota(1),\iota(2)}(\zeta) = \frac{K}{2}\operatorname{Tr}(2^2 M_{\iota(2)}\Sigma M_{\iota(1)}\Sigma) + \frac{1}{2}\sum_{p\in\mathcal{P}(\iota(1),\iota(2))}\operatorname{Tr}(2^2 M_{p(1)}\Sigma M_{p(2)}\Sigma\Omega) \tag{B.18}$$

$$\begin{aligned}\kappa_3^{\iota(1),\iota(2),\iota(3)}(\zeta) =& \frac{K}{2}\sum_{p\in\mathcal{P}(\iota(2),\iota(3))}\operatorname{Tr}(2^3 M_{p(1)}\Sigma M_{p(2)}\Sigma M_{\iota(1)}\Sigma) \\ &+ \frac{1}{2}\sum_{p\in\mathcal{P}(\iota(1),\iota(2),\iota(3))}\operatorname{Tr}(2^3 M_{p(1)}\Sigma M_{p(2)}\Sigma M_{p(3)}\Sigma\Omega)\end{aligned} \tag{B.19}$$

$$\begin{aligned}\kappa_n^{\iota(1),\iota(2),\cdots}(\zeta) =& 2^{n-1}\left(K\sum_{p\in\mathcal{P}(\iota(2),\iota(3),\ldots)}\operatorname{Tr}(M_{p(1)}\Sigma M_{p(2)}\Sigma\ldots M_{p(n-1)}\Sigma M_{\iota(1)}\Sigma)\right. \\ &\left.+\sum_{p\in\mathcal{P}(\iota(1),\iota(2),\iota(3),\ldots)}\operatorname{Tr}(M_{p(1)}\Sigma M_{p(2)}\Sigma M_{p(3)}\Sigma\ldots M_{p(n)}\Sigma\Omega)\right),\end{aligned} \tag{B.20}$$

where $\mathcal{P}(\cdot)$ denotes the set of all permutations of the indices in the argument and $p(i)$ denotes the i-th index in a given permutation. Note that for $N = 1$, the formulas collapse to those for the trace in Mathai (1980).

Denote the row block index of $\iota(i)$ as $\iota(i)_1$ and the column block index as $\iota(i)_2$, so $M_{\iota(i)} = M_{\iota(i)_1,\iota(i)_2}$. Additionally, let $U_{i\bullet}$ denote the i-th block of K rows of the matrix U, and similarly $U_{\bullet i}$ for the block of columns. Lemma **??** in the Online Appendix relates $\operatorname{Tr}(M_{\iota(1)}\Sigma M_{\iota(2)}\Sigma M_{\iota(3)}\Sigma\ldots M_{\iota(n)}\Sigma)$ to $\operatorname{Tr}(\Sigma_{\iota(1)}\Sigma_{\iota(2)}\Sigma_{\iota(3)}\ldots\Sigma_{\iota(n)})$. Applying it to the cumulants above yields

$$\kappa_1^{\iota(1)}(\zeta) = \frac{K}{2}\operatorname{Tr}(2\Sigma_{\iota(1)}) + \frac{1}{2}\operatorname{Tr}(2\Sigma_{\iota(1)}\Omega)$$

$$\kappa_2^{\iota(1),\iota(2)}(\zeta) = \frac{K}{2}\operatorname{Tr}(2^2\Sigma_{\iota(2)_1,\iota(1)_2}\Sigma_{\iota(1)_1,\iota(2)_2}) + \frac{1}{2}\sum_{p\in\mathcal{P}(\iota(1),\iota(2))}\operatorname{Tr}(2^2\Sigma_{p(1)_1,p(2)_2}\Sigma_{p(2)_1,p(1)_2}\Omega)$$

$$\begin{aligned}\kappa_3^{\iota(1),\iota(2),\iota(3)}(\zeta) =& \frac{K}{2}\sum_{\iota\in\mathcal{P}(\iota(2),\iota(3))}\operatorname{Tr}(2^3\Sigma_{p(1)_1,p(2)_2}\Sigma_{p(2)_1,\iota(1)_2}\Sigma_{\iota(1)_1,p(1)_2}) \\ &+ \frac{1}{2}\sum_{\iota\in\mathcal{P}(\iota(1),\iota(2),\iota(3))}\operatorname{Tr}(2^3\Sigma_{p(1)_1,p(2)_2}\Sigma_{p(2)_1p(3)_2}\Sigma_{p(3)_1,p(1)_2}\Omega)\end{aligned}$$

$$\kappa_n^{\iota(1),\iota(2),\cdots}(\zeta) = 2^{n-1}\left(K\sum_{p\in\mathcal{P}(\iota(2),\iota(3),\ldots)}\operatorname{Tr}(\Sigma_{p(1)_1,p(2)_2}\Sigma_{p(2)_1,p(3)_2}\ldots\Sigma_{p(n-1)_1,\iota(1)_2}\Sigma_{\iota(1)_1,p(1)_2})\right.$$

(B.21)

$$+ \sum_{p \in \mathcal{P}(\iota(1),\iota(2),\ldots)} \operatorname{Tr}(\Sigma_{p(1)_1,p(2)_2}\Sigma_{p(2)_1,p(3)_2}\ldots\Sigma_{p(n)_1,p(1)_2}\Omega)\Bigg).$$

We ultimately need the cumulants of $\gamma' R'_{N,K}(\zeta \otimes I_K)R_{N,K}\gamma$. Using the preceding expressions for cumulants of $R'_{N,K}(\zeta \otimes I_K)R_{N,K}$, we can compute the cumulants of such quadratic forms. By homogeneity, the n-th cumulant of the quadratic form $\gamma' U\gamma$ for a random matrix U is given by

$$\kappa_n(\gamma' U\gamma) = \sum_{\iota(1)_1=1}^{N}\sum_{\iota(1)_2=1}^{N}\ldots\sum_{\iota(n)_1=1}^{N}\sum_{\iota(n)_2=1}^{N}\left(\prod_{j=1}^{n}\gamma_{\iota(j)_1}\gamma_{\iota(j)_2}\right)\kappa_n^{\iota(1),\ldots,\iota(n)}(U), \tag{B.22}$$

where $\iota(i)$ denote indices of individual elements of U. Given the previously derived expressions for the cumulants of the entries of $R'_{N,K}(\zeta \otimes I_K)R_{N,K}$, we can now compute the cumulants of $\gamma' R'_{N,K}(\zeta \otimes I_K)R_{N,K}\gamma$, noting that the cumulants for the ij entry are equal to those for the trace of the ij $K \times K$ block of ζ, $\operatorname{Tr}(\zeta_{ij})$.

Plugging in the first summation in $\kappa_n^{\iota(1),\ldots\iota(n)}(\zeta)$, equation (B.21),

$$\begin{aligned}
&\sum_{\iota(1)_1=1}^{N}\sum_{\iota(1)_2=1}^{N}\ldots\sum_{\iota(n)_1=1}^{N}\sum_{\iota(n)_2=1}^{N}\left(\prod_{j=1}^{n}\gamma_{\iota(j)_1}\gamma_{\iota(j)_2}\right)\sum_{p\in\mathcal{P}(\iota(2),\iota(3),\ldots)}\operatorname{Tr}\left(\Sigma_{p(1)_1,p(2)_2}\Sigma_{p(2)_1,p(3)_2}\ldots\right.\\
&\qquad\left.\Sigma_{p(n-1)_1,\iota(1)_2}\Sigma_{\iota(1)_1,p(1)_2}\right)\\
=&\sum_{p\in\mathcal{P}(\iota(2),\iota(3),\ldots)}\sum_{\iota'(1)_1=1}^{N}\sum_{\iota'(1)_2=1}^{N}\ldots\sum_{\iota'(n)_1=1}^{N}\sum_{\iota'(n)_2=1}^{N}\left(\prod_{j=1}^{n}\gamma_{\iota'(j)_1}\gamma_{\iota'(j)_2}\right)\operatorname{Tr}(\Sigma_{\iota'(1)}\ldots\Sigma_{\iota'(n)})
\end{aligned}$$

(B.23)

$$=(n-1)!\sum_{\iota'(1)_1=1}^{N}\sum_{\iota'(1)_2=1}^{N}\ldots\sum_{\iota'(n)_1=1}^{N}\sum_{\iota'(n)_2=1}^{N}\left(\prod_{j=1}^{n}\gamma_{\iota'(j)_1}\gamma_{\iota'(j)_2}\right)\operatorname{Tr}(\Sigma_{\iota'(1)}\ldots\Sigma_{\iota(n)'}),$$

where we used a change of indices to move from the first line to the second (recognizing that each set of permuted indices on the blocks of Σ is just the index for some other block of Σ indexed by $\iota'(i)$) and in moving to the third observed that the summand of the outer summation does not depend on the indices of that summation. By the definition of matrix multiplication and considerable algebra,

$$\operatorname{Tr}(((\gamma\gamma'\otimes I_K)U)^n) = \sum_{\iota(1)_1=1}^{N}\sum_{\iota(1)_2=1}^{N}\ldots\sum_{\iota(n)_1=1}^{N}\sum_{\iota(n)_2=1}^{N}\left(\prod_{j=1}^{n}\gamma_{\iota(j)_1}\gamma_{\iota(j)_2}\right)\operatorname{Tr}(U_{\iota(1)}\ldots U_{\iota(n)}).$$

Thus, the expression further simplifies to

$$(n-1)!\operatorname{Tr}(((\gamma\gamma'\otimes I_K)\Sigma)^n). \tag{B.24}$$

Applying the same steps to the second summation in the cumulants,

$$n! \operatorname{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n\Omega). \tag{B.25}$$

Combining both terms yields the expression in part (i) of the theorem.

Turning to part (ii), using the fact that for a positive semi-definite matrix V, $|\operatorname{Tr}(UV)| \leq \operatorname{maxeval} U \operatorname{Tr}(V)$, see Fact 8.12.29 in Bernstein (2009), and the fact that $\operatorname{Tr}\left(((\gamma\gamma' \otimes I_K)\Sigma)^n\Omega\right) \geq 0$, we have

$$\begin{aligned} \operatorname{Tr}\left(((\gamma\gamma' \otimes I_K)\Sigma)^n\Omega\right) &\leq \operatorname{maxeval}\{((\gamma\gamma' \otimes I_K)\Sigma)^{n-1}\} \operatorname{Tr}\left((\gamma\gamma' \otimes I_K)\Sigma\Omega\right) \\ &= K\lambda_{\min} \operatorname{maxeval}\{((\gamma\gamma' \otimes I_K)\Sigma)^{n-1}\}, \end{aligned} \tag{B.26}$$

where the last step follows from $\operatorname{Tr}\left((\gamma\gamma' \otimes I_K)\Sigma\Omega\right) = \gamma' R'(\Sigma\Omega \otimes I_K)R\gamma = K\lambda_{\min}$. Next note that

$$\begin{aligned} \operatorname{maxeval}\{((\gamma\gamma' \otimes I_K)\Sigma)^{n-1}\} &= (\operatorname{maxeval}\{(\gamma\gamma' \otimes I_K)\Sigma\})^{n-1} \\ &= \left(\operatorname{maxeval}\{\Sigma^{\frac{1}{2}}(\gamma\gamma' \otimes I_K)\Sigma^{\frac{1}{2}}\}\right)^{n-1} \\ &\leq (\operatorname{maxeval}\{\Sigma\} \operatorname{maxeval}\{(\gamma\gamma' \otimes I_K)\})^{n-1} \\ &= \operatorname{maxeval}\{\Sigma\}^{n-1}, \end{aligned} \tag{B.27}$$

where the inequality follows from Ostrowski's theorem, for example, Theorem 4.5.9 in Horn and Johnson (2013), and the last step is due to the fact that the matrix $\gamma\gamma'$ has only one non-zero eigenvalue that is equal to one. We therefore have the inequality

$$\operatorname{Tr}\left(((\gamma\gamma' \otimes I_K)\Sigma)^n\Omega\right) \leq K\lambda_{\min} \operatorname{maxeval}\{\Sigma\}^{n-1}. \tag{B.28}$$

Using the Lieb-Thirring inequality for positive semi-definite matrices, see Bernstein (2009) Fact 8.12.17,

$$\begin{aligned} \operatorname{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n) &\leq \operatorname{Tr}((\gamma\gamma' \otimes I_K)^n\Sigma^n) \\ &= \operatorname{Tr}(((\gamma\gamma')^n \otimes I_K)\Sigma^n) = \operatorname{Tr}((\gamma\gamma' \otimes I_K)\Sigma^n), \end{aligned} \tag{B.29}$$

where the last equality results from the fact that the matrix $\gamma\gamma'$ has only one non-zero eigenvalue that is equal to one.

Since $\operatorname{Tr}((\gamma\gamma' \otimes I_K)\Sigma^n) = \gamma' R'(\Sigma^n \otimes I_K)R\gamma$ with $\gamma'\gamma = 1$, we have

$$\operatorname{Tr}(((\gamma\gamma' \otimes I_K)\Sigma)^n) \leq \operatorname{maxeval}\{R'(\Sigma^n \otimes I_K)R\}. \tag{B.30}$$

Applying the inequalities (B.28) and (B.30) yields the bounds in part (ii).