Supplementary Appendix to:

Valid Two-Step Identification-Robust Confidence Sets for GMM

Isaiah Andrews

First Version, December 27, 2013. This Version, March 9, 2017

This supplementary appendix contains the proof of an auxiliary result used for the proofs in the main text, as well as a description of the data and additional results for our empirical application, and simulation results for the linear instrumental variables model.

Appendix A: Additional Proofs

Lemma 3 *Under Assumption 5, $\hat{\theta} \to_p \theta_0$ under all $\xi_0 \in \Xi_S$.*

Proof of Lemma 3 The proof is standard but is included for completeness. By Assumption 5(1), (2), and (3), $g_T(\theta_0)' \hat{\Omega} g_T(\theta_0) \to_p 0$. Consider any $\varepsilon > 0$ and note that by Assumption 5(4),

$$\inf_{||\theta' - \theta_0|| \geq \varepsilon} \left(\lim_{T\to\infty} E_{T,\xi_0}[g_t(\theta)] \right)' \Omega(\theta) \left(\lim_{T\to\infty} E_{T,\xi_0}[g_t(\theta)] \right) > \delta$$

for some $\delta > 0$. By Assumption 5(3) the eigenvalues of $\Omega(\theta)$ are uniformly bounded both above and away from zero. Thus, by the Continuous Mapping Theorem

$$Pr_{T,\xi_0}\left\{ \sup_{\theta\in\Theta} \left| \left(\lim_{T\to\infty} E_{T,\xi_0}[g_t(\theta)] \right)' \Omega(\theta) \left(\lim_{T\to\infty} E_{T,\xi_0}[g_t(\theta)] \right) - g_T(\theta)' \hat{\Omega}(\theta) g_T(\theta) \right| > \frac{\delta}{3} \right\} \to 0.$$

This in turn implies that

$$Pr_{T,\xi_0}\left\{ \inf_{||\theta' - \theta_0|| \geq \varepsilon} g_T(\theta)' \hat{\Omega}(\theta) g_T(\theta) < \frac{2\delta}{3} \right\} \to 0$$

while by point-wise convergence

$$Pr_{T,\xi_0}\left\{g_T(\theta_0)'\hat{\Omega}(\theta_0)\,g_T(\theta_0) > \frac{\delta}{3}\right\} \to 0.$$

Hence, $Pr_{T,\xi_0}\left\{||\hat{\theta}-\theta_0|| \geq \varepsilon\right\} \to 0$. Since the same argument holds for any $\varepsilon > 0$, consistency follows immediately.□

Proof of Lemma 2

The proof is standard, but is included for completeness. If $A_{\theta,T}$ is empty I have defined $\sup_{\theta\in A_{\theta,T}}\left\|W\left(f\left(\theta\right)\right)-K_{\Omega,}$ 0. Hence, I restrict attention to non-empty realizations of $A_{\theta,T}$ and condition on $A_{\theta,T} \neq \emptyset$ for the remainder of the analysis. We know that for $B(\theta_0)$ as in Assumption 6, $Pr\left\{A_{\theta,T} \subset B(\theta_0)\right\} \to 1$. Hence, by the assumptions, $\sup_{\theta\in A_{\theta,T}}\|G_T(\theta)-J(\theta_0)\| = o_p(1)$, $\sup_{\theta\in A_{\theta,T}}\|D_T(\theta)-J(\theta_0)\| = o_p(1)$, and $\sup_{\theta\in A_{\theta,T}}\left\|\hat{\Sigma}_g(\theta)-\Sigma_g(\theta_0)\right\| = o_p(1)$. By a mean value expansion

$$\begin{gathered}\left(D_T(\theta)'\hat{\Omega}(\theta)\,D_T(\theta)\right)^{-1}D_T(\theta)'\hat{\Omega}(\theta)\,g_T(\theta)\\ = \left(D_T(\theta)'\hat{\Omega}(\theta)\,D_T(\theta)\right)^{-1}D_T(\theta)'\,\hat{\Omega}(\theta)\left(g_T(\theta_0)+G_T(\theta^*)(\theta-\theta_0)\right)\end{gathered}$$

for θ^* a value between θ and θ_0 which can vary across rows. By the Continuous Mapping Theorem and the fact that $J(\theta)$ and $\Omega(\theta)$ are continuous and full rank at θ_0:

$$\begin{gathered}\sup_{\theta\in A_{\theta,T}}\left\|\left(D_T(\theta)'\hat{\Omega}(\theta)\,D_T(\theta)\right)^{-1}D_T(\theta)'\,\hat{\Omega}(\theta)\right.\\ \left.-\left(J(\theta_0)'\Omega(\theta_0)\,J(\theta_0)\right)^{-1}J(\theta_0)'\Omega(\theta_0)\right\| = o_p(1)\end{gathered}$$

while

$$\sup_{\theta\in A_{\theta,T}}\left\|\left(D_T(\theta)'\hat{\Omega}(\theta)\,D_T(\theta)\right)^{-1}D_T(\theta)'\,\hat{\Omega}(\theta)\,G_T(\theta^*)-I\right\| = o_p(1).$$

Hence,

$$\begin{gathered}\sup_{\theta\in A_{\theta,T}}\left\|\left(D_T(\theta)'\hat{\Omega}(\theta)\,D_T(\theta)\right)^{-1}D_T(\theta)'\,\hat{\Omega}(\theta)\,g_T(\theta)\right.\\ \left.-\left(J(\theta_0)'\Omega(\theta_0)\,J(\theta_0)\right)^{-1}J(\theta_0)'\Omega(\theta_0)\,g_T(\theta_0)-(\theta-\theta_0)\right\| = o_p\left(\frac{1}{\sqrt{T}}\right).\end{gathered}$$

Similar arguments, together with Lemma 3, establish that

$$\left\| \left(J(\theta_0)'\Omega(\theta_0)J(\theta_0)\right)^{-1} J(\theta_0)'\Omega(\theta_0)\,g_T(\theta_0) + \left(\hat{\theta}-\theta_0\right) \right\| = o_p\left(\frac{1}{\sqrt{T}}\right)$$

and hence, by the triangle inequality, that

$$\sup_{\theta\in A_{\theta,T}} \left\| \left(D_T(\theta)'\hat{\Omega}(\theta)D_T(\theta)\right)^{-1} D_T(\theta)'\,\hat{\Omega}(\theta)\,g_T(\theta) - \left(\hat{\theta}-\theta\right) \right\| = o_p\left(\frac{1}{\sqrt{T}}\right).$$

Since

$$\begin{gathered}\sup\nolimits_{\theta\in A_{\theta,T}} \Big\| \left(D_T(\theta)'\hat{\Omega}(\theta)\hat{\Sigma}_g(\theta)\hat{\Omega}(\theta)D_T(\theta)\right)^{-1} \\ - \left(G_T(\hat{\theta})'\hat{\Omega}\left(\hat{\theta}\right)\hat{\Sigma}_g\left(\hat{\theta}\right)\hat{\Omega}\left(\hat{\theta}\right)G_T\left(\hat{\theta}\right)\right)^{-1} \Big\| = o_p(1)\end{gathered}$$

under the assumptions, this suffices to establish the desired equivalence for test statistics of the full parameter vector, which take $f(\theta) = \theta$. To complete the proof, I need only show that the same result holds for general $f(\cdot)$, which follows from Δ-method arguments. In particular, as noted in van der Vaart (2000) Theorem 3.1, under the assumptions

$$\left\| \sqrt{T}\left(f\left(\hat{\theta}\right) - f(\theta_0)\right) - \sqrt{T}\frac{\partial}{\partial\theta'}f(\theta_0)\left(\hat{\theta}-\theta_0\right) \right\| = o_p(1).$$

Since $\sup_{\theta\in A_{\theta,T}} \|\theta-\theta_0\| = o_p(1)$, by the definition of differentiability

$$\sup_{\theta\in A_{\theta,T}} \left\| \frac{f(\theta)-f(\theta_0)-\frac{\partial}{\partial\theta'}f(\theta_0)(\theta-\theta_0)}{\|\theta-\theta_0\|} \right\| = o_p(1)$$

which implies that

$$\sup_{\theta\in A_{\theta,T}} \left\| \sqrt{T}\left(f(\theta) - f(\theta_0)\right) - \sqrt{T}\frac{\partial}{\partial\theta'}f(\theta_0)\left(\theta-\theta_0\right) \right\| = o_p(1)$$

and hence by the triangle inequality $\left\| \sqrt{T}\left(f\left(\hat{\theta}\right) - f(\theta)\right) - \sqrt{T}\frac{\partial}{\partial\theta'}f(\theta_0)\left(\hat{\theta}-\theta\right) \right\| = o_p(1)$, yielding the statement:

$$\left\| \sqrt{T}\left(f\left(\hat{\theta}\right) - f(\theta)\right) + \sqrt{T}\frac{\partial}{\partial\theta'}f(\theta_0)\left(D_T(\theta)'\hat{\Omega}(\theta)D_T(\theta)\right)^{-1} D_T(\theta)'\,\hat{\Omega}(\theta)\,g_T(\theta) \right\| = o_p(1).$$

Since $\sup_{\theta \in A_{T,\theta}} \left\| \frac{\partial}{\partial \theta'} f(\theta_0) - \frac{\partial}{\partial \theta'} f(\theta) \right\| = o_p(1)$ and $\frac{\partial}{\partial \theta'} f(\theta_0)$ is full rank,

$$\begin{aligned} \sup_{\theta \in A_{\theta,T}} \Bigg\| & \Big(\tfrac{\partial}{\partial \theta'} f(\theta) \left(D_T(\theta)' \hat{\Omega}(\theta) D_T(\theta) \right)^{-1} D_T(\theta)' \hat{\Omega}(\theta) \\ & \times \hat{\Sigma}_g(\theta) \hat{\Omega}(\theta) \hat{D}_T(\theta) \left(D_T(\theta)' \hat{\Omega}(\theta) D_T(\theta) \right)^{-1} \tfrac{\partial}{\partial \theta'} f(\theta)' \Big)^{-1} \\ & - \Big(\tfrac{\partial}{\partial \theta'} f(\theta_0) \left(J(\theta_0)' \Omega(\theta_0) J(\theta_0) \right)^{-1} J(\theta_0)' \hat{\Omega}(\theta_0) \hat{\Sigma}_g(\theta_0)^{-1} \\ & \times \hat{\Omega}(\theta_0) J(\theta_0) \left(J(\theta_0)' \Omega(\theta_0) J(\theta_0) \right)^{-1} \tfrac{\partial}{\partial \theta'} f(\theta_0)' \Big)^{-1} \Bigg\| = o_p(1) \end{aligned}$$

by the Continuous Mapping Theorem, and by the triangle inequality

$$\begin{aligned} \sup_{\theta \in A_{\theta,T}} \Bigg\| & \Big(\tfrac{\partial}{\partial \theta'} f(\hat{\theta}) \left(G_T(\hat{\theta})' \hat{\Omega}(\hat{\theta}) G_T(\hat{\theta}) \right)^{-1} G_T(\hat{\theta})' \hat{\Omega}(\hat{\theta})' \\ & \times \hat{\Sigma}_g(\hat{\theta})^{-1} \hat{\Omega}(\hat{\theta}) G_T(\hat{\theta}) \left(G_T(\hat{\theta})' \hat{\Omega}(\hat{\theta}) G_T(\hat{\theta}) \right)^{-1} \tfrac{\partial}{\partial \theta'} f(\hat{\theta})' \Big)^{-1} \\ & - \Big(\tfrac{\partial}{\partial \theta'} f(\theta) \left(D_T(\theta)' \hat{\Omega}(\theta) D_T(\theta) \right)^{-1} D_T(\theta)' \hat{\Omega}(\theta) \hat{\Sigma}_g(\theta)^{-1} \\ & \times \hat{\Omega}(\theta) D_T(\theta) \left(D_T(\theta)' \hat{\Omega}(\theta) D_T(\theta) \right)^{-1} \tfrac{\partial}{\partial \theta'} f(\theta)' \Big)^{-1} \Bigg\| = o_p(1). \end{aligned}$$

Hence $\sup_{\theta \in A_{\theta,T}} \left| W(f(\theta)) - K_{\Omega,f}(\theta) \right| = o_p(1)$, so the $K_{\Omega,f}$ and Wald statistics are first-order equivalent on $A_{\theta,T}$.□

Appendix B: Data Description and Additional Empirical Results

As noted in the main text, for the empirical application I follow Stock and Wright (2000) and use an extension of the long annual dataset of Campbell and Shiller (1987), which consists of annual US data from 1873 to 1993. As discussed in Stock and Wright (2000), the interest rate is the nominal rate for four to six month commercial paper, while stock returns are based on the Cowles Commission index for the first part of the sample, and the annual average price of the S&P monthly composite index for the second part. Asset returns are converted to real terms using the producer price index, and consumption is measured as real consumption of nondurables and services per capita.

To complement the empirical results in the main text, here we discuss results from augmenting the set of moments considered with moments which take R_t to equal the interest rate. As with our treatment of the moments based on the equity return, we instrument these moments with a constant,

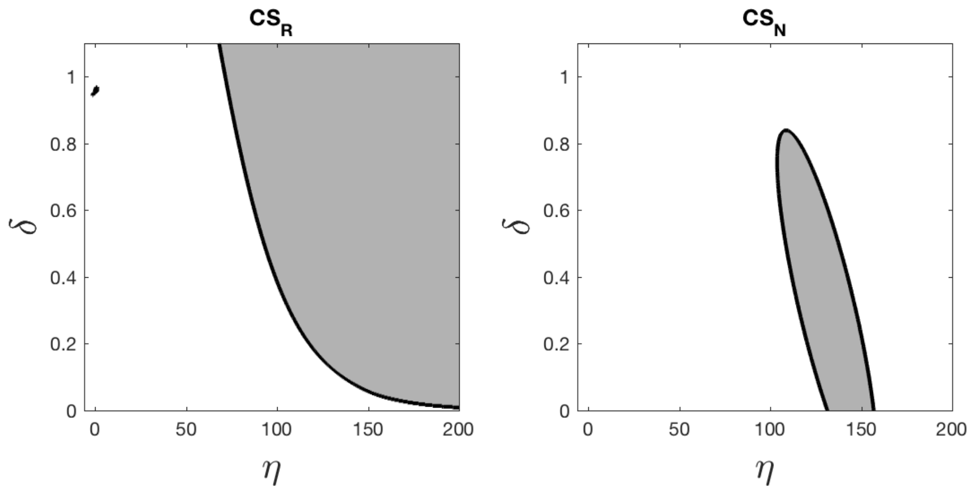

Figure 2: Robust and non-robust confidence sets for full parameter vector θ. The distortion cutoff $\hat{\gamma}$ is equal to 69.99%.

Parameter	CS_R	CS_N	$\hat{\gamma}$
δ	$[.01, 1.1]$	$[0, 0.73]$	65.72%
η	$[-1.15, -0.8] \cup [0.05, 0.95] \cup [56.6, 200]$	$[109.3, 153.1]$	67.49%

Table 2: Confidence sets and distortion cutoffs $\hat{\gamma}$ for parameters δ and η.

lagged consumption growth, and the first lag of the interest rate. Results from these specifications are reported in Figure 2 and Table 2. As we see from these results, estimated risk aversion is much larger with this choice of moments, and the maximal distortion cutoffs $\hat{\gamma}$ are likewise much larger than in our baseline specifications.

Appendix C: IV Simulation Results

To illustrate the performance of the two-step procedure proposed in the main text, this appendix describes simulation results for the linear instrumental variables model. I first consider the simulation performance of two-step procedures based on the first stage F-statistic with Stock and Yogo (2005) (abbreviated SY for the remainder of this appendix) critical values in linear IV with homoskedastic and non-homoskedastic errors, and then turn to the performance of the two-step approach proposed in the main text.

As in the IV example in main text I focus on the linear IV model with a single endogenous regressor, where there either are no additional exogenous regressors or any such regressors have already been partialled out.[17] As before the model, written in reduced form, is

$$Y = Z\pi_0\beta_0 + V_1$$
$$X = Z\pi_0 + V_2$$

for Z a $T \times k$ matrix of instruments, X a $T \times 1$ vector of endogenous regressors, Y a $T \times 1$ vector of outcome variables, and V_1 and V_2 both $T \times 1$ vectors of residuals, where I assume that $E[V_{1,t}Z_t] = E[V_{2,t}Z_t] = 0$ for Z_t the transpose of row t of Z.

I am interested in constructing confidence sets for the scalar coefficient β, treating the $k \times 1$ vector of first-stage parameters π as nuisance parameters. A common nominal level $1-\alpha$ confidence set in empirical practice is the two stage least squares (2SLS) Wald confidence set, equal to the two-stage least squares estimator $\hat{\beta}_{2SLS}$ plus and minus a multiple of the standard error. We can likewise construct Wald confidence sets based on other estimators $\hat{\beta}$, for example limited information maximum likelihood (LIML) or, in the heteroskedastic case, efficient two-step GMM (2SGMM) or continuously updating GMM (CUGMM).

C.1 Two-Step F-Statistic Confidence Sets

All of these Wald confidence sets may exhibit large coverage distortions when the first-stage parameter π_0 is small. The first stage F-statistic aims to measure the magnitude of π and, for $\hat{\pi} = (Z'Z)^{-1} Z'X$ the OLS estimator of π and $\widehat{\Sigma}_{\hat{\pi}}$ an estimator for the variance of $\sqrt{T}(\hat{\pi}-\pi_0)$, is equal to $F = \frac{T}{k}\hat{\pi}'\widehat{\Sigma}_{\hat{\pi}}^{-1}\hat{\pi}$. While the conventional first-stage F-statistic uses an estimator $\hat{\Sigma}_{\hat{\pi}}$ which assumes the errors (V_1, V_2) are conditionally homoskedastic given Z, this can yield highly unreliable results when the errors are in fact heteroskedastic (or serially correlated or clustered). Hence, throughout this section I will focus on the heteroskedasticity-robust F-statistic, which takes $\widehat{\Sigma}_{\hat{\pi}}$ to be the White (1980) covariance matrix estimator, and one can likewise define serial-correlation and clustering robust F-statistics when appropriate.

To construct a two-step confidence set as in (1) I also need to define an appropriate robust confidence set CS_R. For simplicity I consider confidence sets based on the S statistic of Stock and Wright (2000), (8), where $g_T(\beta) = \frac{1}{T}\sum Z_t(Y_t - \beta X_t)$ and $\widehat{\Sigma}_g(\beta)$ is the usual heteroskedasticity-

[17] That is, for exogenous controls W and initial data $(\tilde{Y}, \tilde{X}, \tilde{Z}, W)$, $Y = M_W\tilde{Y}$, $X = M_W\tilde{X}$, $Z = M_W\tilde{Z}$, where $M_W = I - W(W'W)^{-1}W'$.

robust variance estimator for $\sqrt{T}g_T(\beta)$. As shown by Stock and Wright (2000), the S statistic evaluated at the true parameter value β_0 will be approximately χ^2_k distributed in large samples regardless of identification strength, so the level $1-\alpha$ confidence set for β based on this statistic is $CS_S = \left\{\beta : S(\beta) < \chi^2_{k,1-\alpha}\right\}$.

In the rest of this section I focus on two-step confidence sets based on the first stage F-statistic. In particular, I consider CS_2 defined as in (1), with $\phi_{ICS} = 1\{F < c\}$ for some cutoff c, let CS_N be some Wald confidence set, and take $CS_R = CS_S$. I study the coverage of these confidence sets for different choices of Wald confidence set and cutoff. I first highlight that, as expected given the results of SY, the conventional rule of thumb cutoff $c = 10$ does not control coverage for two-step confidence sets even in homoskedastic models. Next, I study cutoffs based on SY, which by their results ensure bounded coverage distortions in homoskedastic models. I provide what is, to my knowledge, the first demonstration in the literature that the SY cutoffs fail to control coverage distortions under heteroskedasticity in over-identified models, even when using the heteroskedasticity-robust F-statistic. While this result is unsurprising given that the results of SY are derived only for models with homoskedastic errors, it raises questions about the widespread application the first-stage F-statistic in empirical contexts with non-homoskedastic errors.

The simulations set $\beta_0 = 0$ and assume that Z_t is a collection of dummy variables for different values of a categorical instrument $\tilde{Z}_t \in \{1,...,k\}$. I take $k \in \{5,10,20\}$ and for each k consider two calibrations: one with a moderate degree of endogeneity and the other with a very high degree of endogeneity. In each calibration I consider a wide range of values for identification strength (as measured by $\|\pi_0\|$, where I hold $\pi_0/\|\pi_0\|$ fixed) ranging from non-identification to very strong identification, and report the smallest coverage probability for each confidence set over these different values, $\min_{\|\pi_0\|} Pr_{\pi_0,\beta_0}\{\beta_0 \in CS\}$. All simulations are based on samples of 10,000 observations.[18] Further details on the simulation design may be found at the end of this appendix.

C.1.1 The First Stage F-Statistic Under Homoskedasticity

I first study the performance of nominal 95% two-step confidence sets based on the F-statistic together with different cutoffs c and Wald confidence sets CS_N under homoskedasticity. I begin

[18] I take the sample size to be large to highlight that the poor performance of two-step confidence sets based on the first-stage F-statistic under heteroskedasticity is not due to small-sample problems with heteroskedasticity-robust covariance matrix estimation, and use this sample size throughout for consistency.

with the usual LIML and 2SLS confidence sets. Next, I consider rule of thumb confidence sets which take $c = 10$. Finally I use cutoffs based on critical values from SY, specifically $c = 26.87$, $c = 38.54$, and $c = 62.30$ for $k = 5$, $k = 10$, and $k = 20$ respectively for CS_N the 2SLS confidence set, and $c = 5.44$, $c = 3.68$, and $c = 3.21$ for CS_N the LIML confidence set.[19] The results of SY imply that in models with homoskedastic errors this choice of cutoffs ensures coverage distortions no larger than 10%, and so coverage no less that 85%, for two-step confidence sets with nominal coverage 95%.[20]

The results of this exercise are reported in Table 3. As expected, the rule of thumb cutoff of 10 does not ensure any fixed level of coverage for two-step confidence sets: while 2SLS confidence sets based on the rule of thumb have coverage distortions less than 10% in the moderate endogeneity calibrations, they exhibit more substantial distortions in the high endogeneity calibrations, and the degree of distortion is increasing in the number of instruments k. By contrast, two-step confidence sets based on the cutoffs of SY have distortions not exceeding 10% (and thus coverage not less than 85%) in all cases, also as expected.

C.1.2 The First Stage F-Statistic Under Heteroskedasticity

I next repeat the simulation exercise taking the errors to be heteroskedastic so $Var\left(\left(V_{1,t}, V_{2,t}\right)|Z_t\right)$ depends on Z_t. Since 2SLS and LIML are inefficient under heteroskedasticity I also consider confidence sets based on CUGMM and 2SGMM. When considering two-step confidence sets based on SY I use LIML cutoffs for CUGMM and 2SLS cutoffs for 2SGMM.[21]

The minimal coverage for all confidence sets considered is reported in Table 4. As in the homoskedastic case neither Wald confidence sets nor two-step confidence sets based on the rule of thumb cutoffs control coverage distortions. Unlike in the homoskedastic case two-step confidence

[19] I obtain these cutoffs by taking $\phi_{ICS} = 1$ when the 5% F-test of SY cannot reject the hypothesis that the nominal 5% Wald test of interest has true size exceeding 10%. This results in a two-step confidence set with coverage at least 85%. While the Wald confidence set has coverage at least 90% under the alternative in the SY pretest, to obtain the coverage of the two-step confidence set we must also account for errors due to the pretest itself. Taking this error into account via a Bonferroni correction gives an 85% lower bound on the coverage of the two step confidence set. As clear from Table 5, this additional correction is necessary (if potentially conservative) as two-step confidence sets based on the SY pretest do not always have 90% coverage.

[20] My focus on maximal 10% distortions is not an endorsement of this choice of γ, but rather due to the fact that this is the smallest γ achievable using the critical values published in SY. One could instead use the approach of SY to derive pretests implying smaller values of γ in the homoskedastic case.

[21] I also considered CUGMM and 2SGMM in the homoskedastic simulations but, unsurprisingly given the large sample size, their behavior is indistinguishable from that of LIML and 2SLS respectively.

	Moderate Endogeneity			High Endogeneity		
Confidence Set	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
LIML CS	57.7%	38.2%	25.2%	11.1%	1.9%	0%
2SLS CS	58.8%	40.4%	42.7%	0%	0%	0%
Rule of thumb LIML CS	92.9%	93.1%	93.5%	90.4%	91.5%	91.4%
Rule of thumb 2SLS CS	90.4%	89.1%	89.2%	82%	76.1%	64.2%
SY LIML CS	91.6%	90.6%	89.4%	88.4%	89.2%	89.7%
SY 2SLS CS	92.6%	92.4%	93.6%	87.5%	87.7%	87.7%

Table 3: Minimal coverage for nominal 95% confidence sets in homoskedastic IV simulations with 10,000 observations, based on 10,000 simulations. LIML CS and 2SLS CS are the usual Wald confidence sets based on LIML and 2SLS, while the rule of thumb confidence sets are two-step confidence sets using the rule-of-thumb cutoff $c = 10$ and the robust S (or Anderson-Rubin) confidence set CS_S. Finally, the SY confidence sets use the SY cutoffs discussed in the text along with the robust S (or Anderson-Rubin) confidence set CS_S, and have asymptotic coverage at least 85% in models with homoskedastic errors.

sets using the SY cutoffs also fail to control coverage distortions, for both efficient and inefficient estimators. More generally, in many cases heteroskedasticity gives rise to far more pronounced coverage shortfalls than those under homoskedasticity.

The central problem with two-step confidence sets based on the SY cutoffs under heteroskedasticity is that the first stage F-statistic used with existing cutoffs is no longer a reliable indicator of identification strength. Related to this finding, Antoine and Renault (2015) derive a measure for identification that is closely related to the first-stage F-statistic under homoskedasticity but may differ substantially under heteroskedasticity. Likewise, the unreliability of conventional F-statistic-based assessments of 2SLS bias under heteroskedasticity was previously highlighted by Bun and de Haan (2010) and Olea and Pfleuger (2013). I find, however, that the issue appears especially stark when considering coverage. In Figure 3 I plot the coverage of Wald confidence sets against the mean of the first-stage F-statistic for the model with ten instruments in the moderate endogeneity calibration as I vary $\|\pi_0\|$, noting that $E[F]$ is a strictly increasing function of $\|\pi_0\|$ in this set-up. As this figure makes clear, even when the mean of the first stage F-statistic is 500, many nominal 95% Wald confidence sets exhibit coverage distortions exceeding 15%.[22] A still more extreme version of this issue arises in the high endogeneity calibration, where the 2SLS confidence set has a 15% coverage distortion even when the mean of the first stage F-statistic is

[22]While the behavior of non-robust confidence sets is similarly poor under the other calibrations, the particular ranking across confidence sets (i.e. which Wald confidence set exhibits the largest distortion) varies.

	Moderate Endogeneity			High Endogeneity		
Confidence Set	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
LIML CS	57.2%	41.3%	27.2%	6.5%	9.5%	1%
2SLS CS	38.3%	27.4%	37.2%	0%	0%	0%
CUGMM CS	28.2%	13%	31.7%	4.8%	0.5%	0%
2SGMM CS	20.6%	9.3%	30.2%	0%	0%	0%
Rule of thumb LIML CS	63.2%	44.8%	31.9%	21.8%	9.5%	1%
Rule of thumb 2SLS CS	55.1%	30.8%	41.8%	0%	0%	0%
Rule of thumb CUGMM CS	45.4%	18.4%	37.3%	71.5%	54.7%	13.1%
Rule of thumb 2SGMM CS	35.4%	13%	34.6%	1.9%	0%	0%
SY LIML CS	61.2%	43%	29.1%	21.8%	9.5%	1%
SY 2SLS CS	63.6%	40.2%	56.1%	0%	0%	0%
SY CUGMM CS	39.5%	15.3%	34.1%	63.2%	54.7%	3.1%
SY 2SGMM CS	46.7%	19.8%	49.2%	2.8%	0%	0%

Table 4: Minimal coverage for nominal 95% confidence sets in heteroskedastic IV simulations with 10,000 observations, based on 10,000 simulations. LIML CS, 2SLS CS, CUGMM CS, and 2SGMM CS are the usual Wald confidence sets based on LIML, 2SLS, CUGMM, and 2SGMM, while the rule of thumb confidence sets are two-step confidence sets using the rule-of-thumb cutoff $c = 10$ and the robust S (or Anderson-Rubin) confidence set CS_S. Finally, the SY confidence sets use the SY cutoffs together with the robust S confidence set CS_S, and have asymptotic coverage at least 85% in models with homoskedastic errors.

100,000. Given these large distortions, it is unsurprising that two-step confidence sets based on the first-stage F-statistic and known cutoffs fail to generate reliable two-step confidence sets in models with heteroskedastic data.

Given these results, it is natural to ask whether some alternative cutoff, or alternative definition of the F-statistic, might render this statistic more appropriate for judging identification strength in over-identified heteroskedastic models. Olea and Pfleuger (2013) show that for the purpose of controlling approximate bias, the answer is affirmative. Unfortunately, however, I have not succeeded in obtaining an analogous result for controlling coverage. Direct extension of the results of SY is far from straightforward, since SY use special structure of the IV model in the homoskedastic case to substantially reduce the dimension of the parameter space, and derive their results by finding the least-favorable error covariance structure. By contrast, in over-identified non-homoskedastic models the simplifications used by SY do not apply, and finding the least-favorable covariance structure appears daunting. Interestingly, in just-identified models with a single endogenous regressor one can show that the structure of IV problem in the non-homoskedastic case continues to

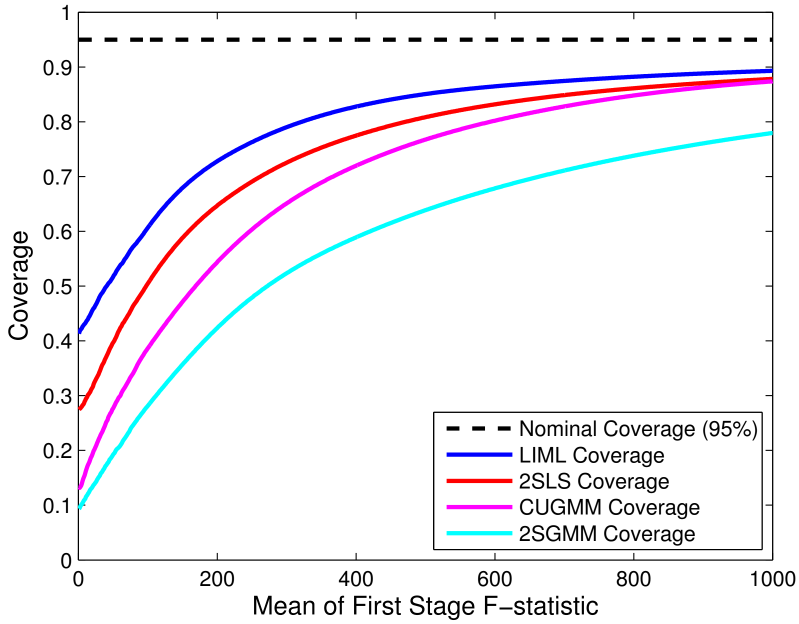

Figure 3: Coverage of Wald confidence sets plotted against the mean of the first stage F-statistic as $\|\pi_0\|$ varies in the moderate endogeneity heteroskedastic linear IV calibration with $k = 10$.

be the same as that studied by SY, with the consequence that their results continue to apply in that case provided we use the heteroskedasticity (or clustering or serial-correlation) robust F-statistic.

C.2 Performance of $CS_2(\gamma)$ in Linear IV

The previous section simulated the performance of two-step confidence sets based on the first-stage F-statistic, and illustrated that while such confidence have well-controlled coverage distortions in models with homoskedastic errors they can have large distortions in models with heteroskedastic errors. In this section, I study the performance of the two-step confidence sets $CS_2(\gamma)$ suggested in the main text in the same simulation designs and find that, as suggested by the theoretical results in the main text, these confidence sets appear to have well-controlled coverage distortions.

C.2.1 Simulation Performance Under Homoskedasticity

I first return to the homoskedastic IV and simulate the coverage of the robust confidence sets CS_R and $CS_P(10\%)$, as well as the two-step confidence sets $CS_2(10\%)$. For comparability with the earlier results, in all cases I set $\alpha = 5\%$ and $\gamma = 10\%$ and continue to consider samples of

	Moderate Endogeneity			High Endogeneity		
Confidence Set	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
CS_R	95%	94.6%	94.7%	94.8%	94.7%	94.6%
$CS_P(10\%)$	84.7%	85%	85.2%	85.4%	85.6%	84.3%
CS_2 LIML	92.6%	92.4%	90.4%	86%	87%	85%
$CS_2(10\%)$ 2SLS	92.8%	92.8%	92.9%	86%	87%	85%

Table 5: Minimal coverage for confidence sets in homoskedastic IV simulations with 10,000 observations, based on 2,500 simulations. CS_R and $CS_P(10\%)$ are robust 95% and 85% confidence sets, respectively, calculated as suggested in Section 3 for $\alpha=5\%$ and $\gamma=10\%$ based on the two-stage least squares weight $\hat{\Omega}(\theta)=\left(\frac{1}{T}Z'Z\right)^{-1}$. CS_2 LIML and CS_2 2SLS are two-step confidence sets (1) based on LIML and 2SLS, calculated as described in Section 3 for $\alpha=5\%$ and $\gamma=10\%$.

10,000 observations. By construction the robust confidence set CS_R has coverage 95% under both weak and strong identification, while $CS_P(10\%)$ has sequential coverage 85% and the two step confidence sets have sequential coverage at least 85%. The results, reported in Table 5, show that the simulated coverage of CS_R and $CS_P(10\%)$ is quite close to their theoretical coverage, while the minimal coverage of the two-step confidence sets is at least 85% in all cases.

Since pretests based on the SY critical values also guarantee coverage at least 85% under homoskedasticity, it is interesting to compare their behavior to that of my ICS statistic (5). In Figure 4 I plot the mean of my ICS statistic $E[\phi_{ICS}(10\%)]$ together with $E\left[\phi_{ICS,SY}\right]$, the mean of the ICS statistic based on the first stage F-statistic with SY's critical values, against the mean of the first-stage F-statistic as I vary $\|\pi_0\|$ in the moderate endogeneity calibration with $k=10$. As we can see, my ICS procedure for LIML behaves quite similarly to that of SY, while my ICS procedure for 2SLS indicates strong identification with substantially higher probability than that of SY. Repeating this exercise for the other moderate endogeneity calibrations I find similar results (not shown), while when I consider the high endogeneity calibrations I find no general ordering between my ICS procedure and that of SY.

C.2.2 Two Step Confidence Sets Under Heteroskedasticity

Let us now return to the heteroskedastic linear IV calibrations. In particular, I consider the robust confidence sets CS_R and $CS_P(10\%)$ based on both the inefficient 2SLS weight matrix $\hat{\Omega}(\theta)=\left(\frac{1}{T}Z'Z\right)^{-1}$ and the efficient weight matrix $\hat{\Omega}(\theta)=\hat{\Sigma}_g(\theta)^{-1}$, as well as two-step confidence sets based on LIML, 2SLS, CUGMM, and 2SGMM. I again take $\alpha=5\%$ and $\gamma=10\%$ in all cases.

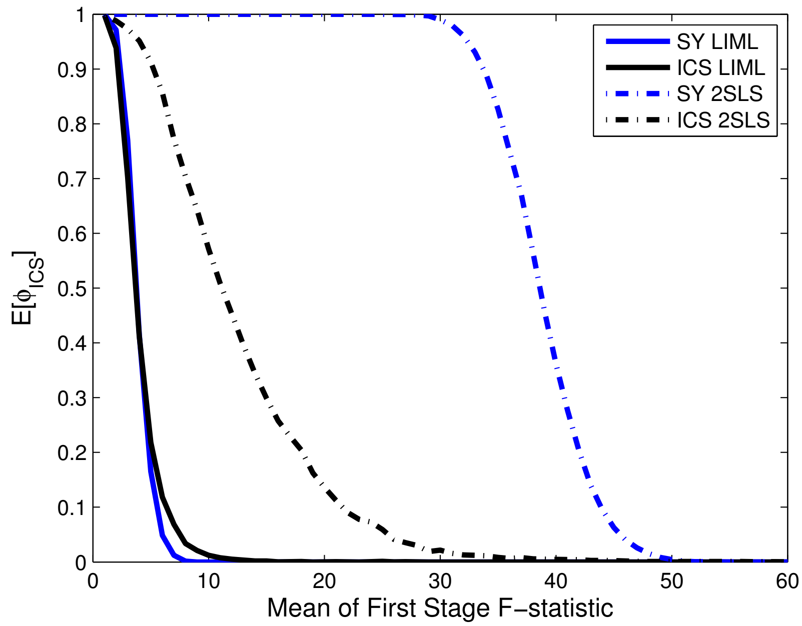

Figure 4: $E[\phi_{ICS}] = Pr\{\phi_{ICS} = 1\}$ plotted against the mean of the first stage F-statistic as $\|\pi_0\|$ varies in the heteroskedastic linear IV calibration with moderate endogeneity and $k = 10$, where SY LIML and SY 2SLS denote pretests based on the first stage F-statistic and the critical values of Stock and Yogo (2005), while ICS LIML and ICS 2SLS use the ICS statistic $\phi_{ICS}(10\%)$ with $\alpha = 5\%$ and $\gamma = 10\%$.

We can see that as in the homoskedastic case my robust confidence sets CS_R and $CS_P(10\%)$ have minimal coverage quite close to their theoretical coverage of 95% and 85%, respectively. Unlike the procedures based on the first stage F-statistic, the confidence sets $CS_2(10\%)$ have minimal coverage at least 85%.

C.3 IV Simulation Design

Heteroskedastic Case To examine the behavior of two-step confidence sets in simulation I need to specify the process generating (Z, V_1, V_2). My focus is on heteroskedasticity, so I consider models where $(Z_t, V_{1,t}, V_{2,t})$ are independent across t but where $Var\left(V_{1,t}, V_{2,t} \middle| Z_t\right)$ may depend on Z_t. I consider a categorical instrument and let Z_t be a collection of dummy variables for different values of $\tilde{Z}_t \in \{1, ..., k\}$, so $Z_t \in \{e_1, ..., e_k\}$ where e_i is the $k \times 1$ vector with 1 in the ith entry and zeros everywhere else. I take $\tilde{Z}_t$ to be uniformly distributed so that $Pr\{Z_t = e_i\} = \frac{1}{k}$ for all $i \in \{1, ..., k\}$ and set the true parameter value $\beta_0 = 0$.

	Moderate Endogeneity			High Endogeneity		
Confidence Set	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
CS_R Inefficient	95.1%	94.7%	95.2%	94.7%	95.2%	94.6%
$CS_P\,(10\%)$ Inefficient	84.7%	84.3%	85%	86%	85%	84.6%
CS_R Efficient	95.1%	94.6%	94.7%	94.9%	94.5%	95%
$CS_P\,(10\%)$ Efficient	84.9%	84.5%	84.6%	84.8%	85.4%	84%
$CS_2\,(10\%)$ LIML	94.1%	92.4%	93%	86.8%	86.8%	85.2%
$CS_2\,(10\%)$ 2SLS	93.7%	94%	94.3%	86.7%	86.6%	85.2%
$CS_2\,(10\%)$ CUGMM	95%	94.1%	93.4%	86.8%	92.6%	88.8%
$CS_2\,(10\%)$ 2SGMM	94.5%	94%	93.9%	86.8%	92.8%	88.8%

Table 6: Minimal coverage for confidence sets in heteroskedastic IV simulations with 10,000 observations, based on 2,500 simulations. CS_R and $CS_P\,(10\%)$ Inefficient are robust 95% and 85% confidence sets (12) based on the two-stage least squares weight matrix $\hat{\Omega}(\theta) = \left(\frac{1}{T}Z'Z\right)^{-1}$, calculated as discussed in the supplement for $\alpha = 5\%$ and $\gamma = 10\%$. CS_R and $CS_P\,(10\%)$ Efficient are robust confidence sets with $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$ calculated as suggested in the main text for $\alpha = 5\%$ and $\gamma = 10\%$. $CS_2\,(10\%)$ LIML, $CS_2\,(10\%)$ 2SLS, $CS_2\,(10\%)$ CUGMM, and $CS_2\,(10\%)$ 2SGMM are two-step confidence sets (1) based on LIML, 2SLS, CUGMM, and 2SGMM, calculated as described in Section 3 for $\alpha = 5\%$ and $\gamma = 10\%$.

Since the support of Z_t is finite I model $Var\left(V_{1,t}, V_{2,t} \middle| Z_t\right)$ fully flexibly and take

$$\begin{pmatrix} V_{1,t} \\ V_{2,t} \end{pmatrix} | Z_t \sim N\left(0, \Sigma_V\left(Z_t\right)\right).$$

To explore the behavior of the model for different parameter values I randomly drew many values of $\Sigma_V\left(Z_t\right)$ and the direction of the first stage $\pi_0/\|\pi_0\|$. For each draw I considered a large range of values for $||\pi_0||$, ranging from non-identification to very strong identification, and for the simulations I focus on particular draws of $\Sigma_V\left(Z_t\right)$ and $\pi_0/\|\pi_0\|$ that generate large coverage distortions for some values of $\|\pi_0\|$. The full specification of $\Sigma_V\left(Z_t\right)$, as well as the accompanying $\pi_0/\|\pi_0\|$, in these designs are reported at the end of this section.

I study models with five, ten, and twenty instruments ($k \in \{5, 10, 20\}$) and in each case consider two calibrations, one with a very high degree of endogeneity as measured by the correlation between the errors $V_{1,t}$ and $V_{2,t}$, and the other with more moderate endogeneity. The space of possible covariance matrices is large, so there likely exist parameter values generating much more pathological behavior for non-robust procedures than I report here. Consequently, my results give only lower bounds for possible coverage distortions. In all cases I simulate samples of 10,000 observations.

To give a sense of the parameter values used in the simulations, in Table 7 I report the (unconditional) correlation between $V_{1,t}$ and $V_{2,t}$ as well as
$Stdev\left(Stdev\left(V_{i,t}|Z_t\right)\right)/Stdev\left(V_{i,t}\right)$, which is a natural measure for the degree of heteroskedasticity.

	Moderate Endogeneity			High Endogeneity		
	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
$Corr\left(V_{1,t},V_{2,t}\right)$	-0.66	-0.59	-0.44	-1.00	1.00	-1.00
$Stdev\left(Stdev\left(V_{1,t}\|Z_t\right)\right)/Stdev\left(V_{1,t}\right)$	0.40	0.53	0.44	0.56	0.55	0.51
$Stdev\left(Stdev\left(V_{2,t}\|Z_t\right)\right)/Stdev\left(V_{2,t}\right)$	0.63	0.55	0.50	0.56	0.55	0.51

Table 7: Summary of linear IV calibration values. Note that $Corr\left(V_{1,t},V_{2,t}\right)$ is in all cases strictly less than one in absolute value, but the reported value is rounded to nearest 0.01.

Homoskedastic Case

For the homoskedastic case, I consider the same simulation calibrations described above, except that in each case I eliminate heteroskedasticity by taking $V_{1,t}$, $V_{2,t}$ to be independent of Z_t with

$$\begin{pmatrix} V_{1,t} \\ V_{2,t} \end{pmatrix} \sim N\left(0, E\left[\Sigma_V\left(Z_t\right)\right]\right).$$

Full Specification of $\Sigma_V(z)$ For Heteroskedastic Case

Finally, Tables 8-13 give the full specification of $\Sigma_V\left(Z_t\right)$ for models with heteroskedastic errors. In particular, recall that Z_t is a vector of dummies generated from a categorical variable $\tilde{Z}_t$. Hence, here I report

$$\Sigma_V\left(\tilde{z}\right) = \begin{pmatrix} Var\left(V_{1,t}|Z_t = e_{\tilde{z}}\right) & Cov\left(V_{1,t},V_{2,t}|Z_t = e_{\tilde{z}}\right) \\ Cov\left(V_{1,t},V_{2,t}|Z_t = e_{\tilde{z}}\right) & Var\left(V_{2,t}|Z_t = e_{\tilde{z}}\right) \end{pmatrix}.$$

$\tilde{z}$	$Var(V_{1,t}\|Z_t = e_{\tilde{z}})$	$Cov(V_{1,t}, V_{2,t}\|Z_t = e_{\tilde{z}})$	$Var(V_{2,t}\|Z_t = e_{\tilde{z}})$	$\pi_0/\\|\pi_0\\|$
1	0.866	-0.528	1.273	0.616
2	3.365	-4.371	5.725	0.45
3	0.653	-0.512	0.414	0.021
4	2.95	-2.9	4.925	-0.327
5	5.84	-0.244	0.01	-0.558

Table 8: Specification of $\Sigma_V(\tilde{z})$ and $\pi_0/\|\pi_0\|$ in moderate endogeneity calibration with $k = 5$.

$\tilde{z}$	$Var(V_{1,t}\|Z_t = e_{\tilde{z}})$	$Cov(V_{1,t}, V_{2,t}\|Z_t = e_{\tilde{z}})$	$Var(V_{2,t}\|Z_t = e_{\tilde{z}})$	$\pi_0/\\|\pi_0\\|$
1	1.102	0.004	0.004	0.57
2	1.891	-2.045	2.789	-0.223
3	3.844	-3.969	4.264	0.478
4	3.457	-1.369	0.779	0.146
5	0.043	0.129	0.395	0.213
6	0.832	-2.383	7.026	0.08
7	1.111	-1.135	1.226	-0.167
8	2.254	-0.39	0.308	0.107
9	10.786	-2.859	1.709	-0.356
10	0.419	-0.934	6.099	-0.396

Table 9: Specification of $\Sigma_V(\tilde{z})$ and $\pi_0/\|\pi_0\|$ in moderate endogeneity calibration with $k = 10$.

Additional References

Antoine, B. and Renault, E. (2015), Testing identification strength. Simon Fraser University Department of Economics Working Paper.

Bun, M. and de Haan, M. (2010), Weak instruments and the first-stage f-statistic in iv modes with a nonscalar error covariance structure. Unpublished Manuscript.

Olea, J. and Pfleuger, C. (2013), ‘A robust test for weak instruments’, *Journal of Business and Economic Statistics* **31**, 358–369.

van der Vaart, A. (2000), *Asymptotic Statistics*, Cambridge University Press.

White, H. (1980), ‘A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity’, *Econometrica* **48**, 817–838.

$\tilde{z}$	$Var(V_{1,t}\|Z_t = e_{\tilde{z}})$	$Cov(V_{1,t}, V_{2,t}\|Z_t = e_{\tilde{z}})$	$Var(V_{2,t}\|Z_t = e_{\tilde{z}})$	$\pi_0/\\|\pi_0\\|$
1	1.904	-1.037	0.791	-0.367
2	1.388	0.645	0.384	-0.166
3	1.391	0.392	0.161	0.184
4	5.13	-4.293	3.671	0.003
5	3.008	0.498	0.141	0.049
6	1.06	-1.499	2.224	0.142
7	1.523	-1.117	1.41	-0.091
8	0.25	-0.134	0.143	0.184
9	5.244	-1.832	0.64	0.34
10	2.4	-3.015	4.918	0.104
11	0.801	0.366	0.381	-0.196
12	1.377	-0.265	0.397	0.524
13	0.198	0.279	0.407	0.004
14	0.385	-0.148	3.186	-0.155
15	3.025	0.044	0.001	-0.289
16	0.055	-0.002	2.286	0.122
17	0.45	-0.415	0.745	0.22
18	0.596	-0.704	2.158	0.16
19	1.115	-0.291	1.731	0.311
20	1.823	-0.779	1.91	-0.079

Table 10: Specification of $\Sigma_V(\tilde{z})$ and $\pi_0/\|\pi_0\|$ in moderate endogeneity calibration with $k = 20$.

$\tilde{z}$	$Var(V_{1,t}\|Z_t = e_{\tilde{z}})$	$Cov(V_{1,t}, V_{2,t}\|Z_t = e_{\tilde{z}})$	$Var(V_{2,t}\|Z_t = e_{\tilde{z}})$	$\pi_0/\\|\pi_0\\|$
1	0.949	-1.13	1.345	0.098
2	0.97	-1.157	1.38	0.08
3	$4.2 \cdot 10^{-5}$	$-5 \cdot 10^{-5}$	$5.8 \cdot 10^{-5}$	-0.778
4	2.671	-3.184	3.796	0.407
5	2.982	-3.556	4.24	0.461

Table 11: Specification of $\Sigma_V(\tilde{z})$ and $\pi_0/\|\pi_0\|$ in high endogeneity calibration with $k = 5$.

$\tilde{z}$	$Var(V_{1,t}\|Z_t = e_{\tilde{z}})$	$Cov(V_{1,t}, V_{2,t}\|Z_t = e_{\tilde{z}})$	$Var(V_{2,t}\|Z_t = e_{\tilde{z}})$	$\pi_0/\\|\pi_0\\|$
1	0.919	1.212	1.6	0.57
2	0.27	0.359	0.477	-0.223
3	1.695	2.245	2.975	0.478
4	0.65	0.861	1.142	0.146
5	0.094	0.128	0.174	0.213
6	0.084	0.11	0.145	0.08
7	2.642	3.508	4.658	-0.167
8	1.113	1.478	1.963	0.107
9	1.708	2.26	2.99	-0.356
10	$2.6\cdot10^{-5}$	$3.2\cdot10^{-5}$	$3.8\cdot10^{-5}$	-0.396

Table 12: Specification of $\Sigma_V(\tilde{z})$ and $\pi_0/\|\pi_0\|$ in high endogeneity calibration with $k = 10$.

$\tilde{z}$	$Var(V_{1,t}\|Z_t = e_{\tilde{z}})$	$Cov(V_{1,t}, V_{2,t}\|Z_t = e_{\tilde{z}})$	$Var(V_{2,t}\|Z_t = e_{\tilde{z}})$	$\pi_0/\\|\pi_0\\|$
1	35.398	-20.896	12.335	-0.367
2	1.197	-0.706	0.417	-0.166
3	$2.6\cdot10^{-4}$	$-1.6\cdot10^{-4}$	$9.8\cdot10^{-5}$	0.184
4	3.979	-2.349	1.387	0.003
5	9.336	-5.512	3.255	0.049
6	7.348	-4.338	2.561	0.142
7	7.202	-4.251	2.51	-0.091
8	0.932	-0.55	0.325	0.184
9	8.089	-4.775	2.818	0.34
10	0.462	-0.273	0.161	0.104
11	9.515	-5.617	3.316	-0.196
12	0.1	-0.059	0.035	0.524
13	2.795	-1.65	0.974	0.004
14	4.927	-2.909	1.717	-0.155
15	7.093	-4.187	2.471	-0.289
16	2.415	-1.426	0.842	0.122
17	4.03	-2.379	1.405	0.22
18	10.161	-5.999	3.542	0.16
19	6.511	-3.843	2.269	0.311
20	5.477	-3.233	1.909	-0.079

Table 13: Specification of $\Sigma_V(\tilde{z})$ and $\pi_0/\|\pi_0\|$ in high endogeneity calibration with $k = 20$.