Massachusetts Institute of Technology
Department of Economics
Working Paper Series

Robust Two-Step Confidence Sets, and the Trouble with the First Stage F-Statistic

Isaiah Andrews

Graduate Student Research Paper 13-01
December 27, 2013

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

Robust Two-Step Confidence Sets, and the Trouble with the First Stage F-Statistic

Isaiah Andrews*

December 27, 2013

Abstract

When weak identification is a concern researchers frequently calculate confidence sets in two steps, first assessing the strength of identification and then, on the basis of this initial assessment, deciding whether to use an identification-robust confidence set. Unfortunately, two-step procedures of this sort can generate highly misleading confidence sets, and we demonstrate that two-step confidence sets based on the first stage F-statistic can have extremely poor coverage in linear instrumental variables models with heteroskedastic errors. To remedy this issue, we introduce a simple approach to detecting weak identification and constructing two-step confidence sets which we show controls coverage distortions under weak identification in general nonlinear GMM models, while also indicating strong identification with probability tending to one if the model is well-identified. Applying our approach to linear IV we show that it is competitive with approaches based on the first-stage F-statistic under homoskedasticity but performs far better under heteroskedasticity.

JEL Classification: C12, C18, C26

Keywords: Confidence Set, Coverage, F-statistic, Pretesting, Weak Identification, Weak Instruments

*Department of Economics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E19-750, Cambridge, MA 02139 USA. Email: iandrews@mit.edu. The author is grateful to Anna Mikusheva, Whitney Newey, and Jerry Hausman for their guidance and support, and to Victor Chernozhukov, Jose Montiel Olea, Nils Wernerfelt, and the participants of the MIT Econometrics Lunch for helpful comments. NSF Graduate Research Fellowship support under grant number 1122374 is gratefully acknowledged.

1 Introduction

In contexts where weak identification is a concern, empirical researchers in economics frequently begin by calculating some statistic intended to measure identification strength. If this preliminary step indicates that identification is not "too" weak, researchers then proceed as usual and calculate non-robust confidence sets, while if weak identification seems to be an issue researchers may calculate identification-robust confidence sets, look for a different specification, or simply decline to report results. Pretests based on the first stage F-statistic in linear instrumental variables (IV) models are the most common example of this practice. Specifically, researchers often calculate the F-statistic for the hypothesis that the model is unidentified and report two-stage least squares confidence sets only if this statistic exceeds the Staiger and Stock (1997) "rule of thumb" cutoff of 10.

We can think of these procedures as constructing a confidence set in two steps, where the first step attempts to asses identification strength and the second step chooses what confidence set to report based on this initial assessment. Unfortunately, unless constructed carefully two-step confidence sets of this sort can exhibit large coverage distortions, in the sense that they may cover the true parameter value with probability substantially smaller than we intend. In this paper we present two main results. First, we show numerically that pretesting procedures based on the first stage F-statistic fail to control coverage distortions in linear IV with heteroskedastic errors, even when we use heteroskedasticity-robust forms of the F-statistic. Second, we propose a novel approach to detecting weak identification which yields two-step confidence sets with bounded coverage distortions under mild conditions. Our approach is computationally straightforward and yields the usual, non-robust confidence sets with probability tending to one under strong identification. Moreover, our results apply to general non-linear GMM models, and so allows us to reliably assess identification strength in a wide range of settings. Simulation results show that our approach performs well in linear IV.

To formally describe two-step confidence sets, let us denote the first-step diagnostic for identification strength, which following D. Andrews and Cheng (2012) we will call an identification category selection (ICS) statistic, by ϕ_{ICS}. We assume that $\phi_{ICS} \in \{0, 1\}$ where

$\phi_{ICS} = 0$ is interpreted as evidence for strong identification while $\phi_{ICS} = 1$ is interpreted as evidence for weak identification. The rule of thumb for the first stage F-statistic, for example, takes ϕ_{ICS} to be an indicator function for $F < 10$. In the second step we use a non-robust confidence set if $\phi_{ICS} = 0$ and a robust confidence set if $\phi_{ICS} = 1$. Denoting the robust and non-robust confidence sets by CS_R and CS_{NR} respectively, this results in the two-step confidence set CS_{2S}

$$CS_{2S} = \begin{cases} CS_{NR} & \text{if } \phi_{ICS} = 0 \\ CS_R & \text{if } \phi_{ICS} = 1 \end{cases}. \tag{1}$$

As noted above, rather than calculating a robust confidence set CS_R when $\phi_{ICS} = 1$, in practice many researchers instead search for a new specification or simply do not report their results. These practices can lead to extremely poor properties for reported confidence sets, so here we focus on "complete" procedures which report a robust confidence set if $\phi_{ICS} = 1$.

Since the intent of researchers using two-step confidence sets is to mitigate inferential problems arising from weak identification, we can ask how successful a given procedure is in achieving this goal. The coverage of a two-step confidence set CS_{2S} can be expressed as

$$\begin{aligned} Pr\left\{\beta_0 \in CS_{2S}\right\} &= Pr\left\{\beta_0 \in CS_{NR} | \phi_{ICS} = 0\right\} \cdot Pr\left\{\phi_{ICS} = 0\right\} \\ &\quad + Pr\left\{\beta_0 \in CS_R | \phi_{ICS} = 1\right\} \cdot Pr\left\{\phi_{ICS} = 1\right\}. \end{aligned}$$

This coverage depends on three terms: the probability that our ICS statistic indicates weak identification, $Pr\left\{\phi_{ICS} = 1\right\}$, the coverage of our non-robust confidence set conditional on $\phi_{ICS} = 0$, $Pr\left\{\beta_0 \in CS_{NR} | \phi_{ICS} = 0\right\}$, and the coverage of our robust confidence set conditional on $\phi_{ICS} = 1$, $Pr\left\{\beta_0 \in CS_R | \phi_{ICS} = 1\right\}$. If we had an ICS statistic which could perfectly detect weak identification, in the sense that $Pr\left\{\phi_{ICS} = 1\right\} = 1$ whenever the coverage of CS_{NR} fell below $1 - \alpha$, then if CS_R had coverage at least $1 - \alpha$ we can see that CS_{2S} would have coverage at least $1 - \alpha$ as well. In practice, however, we must estimate the strength of identification from the data, so our ICS procedures will be imperfect and must be chosen carefully to ensure reasonable performance for two-step procedures. Moreover,

absent a perfect ICS procedure we will typically pay a price for pretesting, often in the form of lower coverage for CS_{2S} than for CS_R at some parameter values. Thus our focus will not be on whether a nominal level $1-\alpha$ confidence set has exactly correct coverage but rather on the extent to which it controls coverage distortions in the sense of having coverage at least $1-\alpha-\gamma$ for some known γ.

As already noted pretests based on the first-stage F-statistic in linear IV are the most commonly used form of ICS procedure. The well-known Staiger and Stock (1997) rule of thumb for linear IV models with a single endogenous regressor declares instruments weak if the first stage F-statistic falls below 10. As discussed by Stock, Wright, and Yogo (2002) this cutoff controls the bias of two stage least squares in linear IV with homoskedastic errors, but unfortunately gives no assurance of controlling coverage for two-step confidence sets. Indeed, as highlighted in the next section two-step confidence sets which take ϕ_{ICS} to be an indicator for the first-stage F-statistic falling below 10 exhibit large coverage distortions in some models. Stock and Yogo (2005) (henceforth SY) show that it *is* possible to construct cutoffs for the first stage F-statistic which control coverage distortions for two-step confidence sets, but these cutoffs will depend on both the estimator used and the number of instruments and, moreover, are derived under the assumption that the data are homoskedastic.

Our first main result, demonstrated numerically in the next section, is that if we allow heteroskedasticity the cutoffs developed by SY no longer control coverage distortions for two-step confidence sets, even when we use heteroskedasticity-robust forms of the F-statistic. Hence, even in the well-studied case of linear IV, commonly-used pretesting procedures fail to control coverage distortions for two-step confidence sets. This is to our knowledge the first demonstration that pretests based on the heteroskedasticity-robust first-stage F-statistic fail to control coverage in heteroskedastic linear IV: while recent work by Bun and de Haan (2010) and Olea and Pflueger (2013) has documented that the first stage F-statistic does not give a reliable guide to the bias of 2SLS in linear IV with non-homoskedastic errors, and Olea and Pflueger (2013) provide an alternative approach to gauge bias in IV models with general covariance structures, neither of these papers addresses questions of testing or confidence set construction.

Our second main result is a procedure for calculating two-step confidence sets in general

GMM models which controls coverage under mild assumptions. These confidence sets are based on a novel ICS statistic ϕ_{ICS} which has the property that $\phi_{ICS} \to_p 0$ under conventional strong-identification asymptotics. Thus under strong identification two-step confidence sets based on our approach coincide with conventional non-robust confidence sets, while under weak identification our two-step confidence sets will control coverage up to an arbitrarily small user-specified maximal distortion.

Alternative Pretests for Weak Identification In addition to tests based on the first stage F-statistic in linear IV, a number of other pretests for identification have been proposed for both linear IV models and more general contexts. One group of procedures, including those of Wright (2003) and Arellano, Hansen, and Sentana (2012), resembles the first stage F-statistic in testing the null hypothesis of identification failure. Since these procedures test the null of underidentification rather than weak identification, if we use them to construct two-step confidence sets their power may be too high. Specifically, they may frequently reject underidentification, and so set $\phi_{ICS} = 0$, even in contexts where non-robust confidence sets are highly unreliable. SY show that one can overcome this problem in linear IV with homoskedastic errors by increasing the critical values for the first stage F-statistic by an appropriate margin but whether a similar adjustment is possible in general models, let alone what an appropriate critical value might be, remains an open question.

Another group of procedures, including those of Hahn and Hausman (2002), Inoue and Rossi (2010), and Wright (2010), takes the opposite approach and tests the null hypothesis of strong identification. For these pretests the concern is that their power may be too low and thus that they may fail to reject strong identification even when non-robust tests are unreliable. In particular, we are unaware of any results guaranteeing that the power of these tests suffices to control coverage distortions for two-step confidence sets under weak identification. To the contrary, Hausman, Stock, and Yogo (2005) document that conventional non-robust confidence sets may exhibit large distortions even when the Hahn-Hausman test has only a low probability of detecting weak instruments, implying that if we base ϕ_{ICS} on the Hahn-Hausman test for weak instruments the resulting two-step confidence set will in some cases exhibit large coverage distortions.

Finally, there are some diagnostics for identification strength proposed in the literature that, together with appropriately chosen robust and non-robust confidence sets, yield two-step confidence sets that control coverage in particular contexts. Specifically, one can view the critical value selection procedures of D. Andrews and Cheng (2012), the pretest-based procedures of Andrews and Mikusheva (2012), and the "switching" procedures of Elliott Mueller and Watson (2012) as particular two-step confidence set procedures where the ICS statistics, together with the robust and non-robust confidence sets, are chosen in such a way that the procedure as a whole controls coverage. Unfortunately, however, these procedures are either limited in their range of application or prohibitively computationally demanding in many cases of economic interest.

In the next section we show numerically that two-step confidence sets based on the first stage F-statistic and known critical values fail to control coverage distortions for two-step confidence sets. Section 3 introduces a robust two-step confidence set and give some results on its properties. This section focuses on the definition, intuition, and implementation of our robust procedures while formal derivation of their asymptotic properties under weak and strong identification is deferred to Sections 5 and 6. Section 4 demonstrates that our approach controls coverage distortions in linear IV with heteroskedasticity and is competitive with F-statistic-based approaches under homoskedasticity. Section 5 shows that our approach can be used to construct robust two-step confidence sets whenever an appropriate robust confidence set is available, and Section 6 shows that our approach is valid in general nonlinear GMM models.

2 The Trouble with the First Stage F-Statistic

Pretests based on the first stage F-statistic are ubiquitous in empirical practice. Unfortunately, two-step confidence sets based on these procedures suffer from a number of difficulties. First, as noted in Stock, Wright, and Yogo (2002) the commonly-used rule of thumb cutoff of 10 does not correspond to any fixed level of coverage control for two-step confidence sets: to bound coverage distortions, one must instead use the cutoffs developed in SY which depend on the number of instruments and the estimator used to construct the non-robust confidence

set CS_{NR}. More worryingly, even the cutoffs derived by SY depend on the assumption of homoskedasticity and have not been shown work outside of the homoskedastic case. Indeed, in this section we demonstrate numerically that two-step confidence sets based on the first stage F-statistic with SY's cutoffs fail to control coverage distortions in linear IV with heteroskedastic errors.

To frame our discussion, we focus on the linear IV model with a single endogenous regressor, where we assume that either there are no additional exogenous regressors or that any such regressors have already been partialled out. 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 we 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.

We are 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

$$CS_{2SLS} = \left[\hat{\beta}_{2SLS} - c_{\alpha/2}\frac{\hat{\sigma}_{2SLS}}{\sqrt{T}}, \hat{\beta}_{2SLS} + c_{\alpha/2}\frac{\hat{\sigma}_{2SLS}}{\sqrt{T}}\right] \tag{2}$$

where $c_{\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution, $\hat{\beta}_{2SLS}$ is the 2SLS estimator for β_0, and $\hat{\sigma}_{2SLS}$ is an estimator of the standard deviation of $\sqrt{T}\hat{\beta}_{2SLS}$. 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 continuous updating GMM (CUGMM). As is now well-known, however, confidence sets based on all these estimators may exhibit large coverage distortions when π is small.

The common approach to assessing whether π is "too" small is based on the first-stage F-statistic for testing the hypothesis that $\pi = 0$. If we let $\hat{\pi}$ be the OLS estimate of π_0,

$\hat{\pi} = (Z'Z)^{-1} Z'X$, the first stage F-statistic is

$$F = \frac{T-k}{k} \hat{\pi}' \hat{\Sigma}_{\hat{\pi}}^{-1} \hat{\pi}$$

for $\hat{\Sigma}_{\pi}$ an estimator for the variance of $\sqrt{T}(\hat{\pi} - \pi_0)$. The conventional first-stage F-statistic, which assumes the errors are homoskedastic, uses

$$\hat{\Sigma}_{\hat{\pi}} = \left(\frac{1}{T-k} \hat{V}_2' \hat{V}_2 - \left(\frac{1}{T-k} \sum_t \hat{V}_{2,t} \right)^2 \right) \frac{1}{T-k} Z'Z$$

where $\hat{V}_2 = Y - Z\hat{\pi}$. Unfortunately, this statistic relies critically on the assumption of homoskedasticity, and may behave erratically if $Var(V_{2,t}|Z_t)$ depends on Z_t. By contrast the heteroskedasticity-robust F-statistic, based on the White (1980) covariance matrix estimator, uses

$$\hat{\Sigma}_{\hat{\pi}} = \frac{1}{T} \sum \hat{V}_{2,t}^2 Z_t Z_t - \left(\frac{1}{T} \sum_t \hat{V}_{2,t} Z_t \right) \left(\frac{1}{T} \sum_t \hat{V}_{2,t} Z_t \right)'.$$

The robust and non-robust F-statistics are asymptotically equivalent and distributed $\frac{1}{k}\chi_k^2$ asymptotically when the data are homoskedastic, but under heteroskedasticity the robust F-statistic continues to have a scaled χ^2 distribution asymptotically while the non-robust F-statistic does not. One can likewise define serial correlation and clustering-robust F-statistics by using appropriate robust estimators $\hat{\Sigma}_{\hat{\pi}}$.[1] Unsurprisingly, two-step procedures based on the non-robust first stage F-statistic are unreliable when used with heteroskedastic data, so going forward we focus on the heteroskedasticity-robust F-statistic.

To construct a two-step confidence set as in (1) we need to define an appropriate robust confidence set. Here we consider confidence sets based on the S statistic of Stock and Wright (2000) (a generalization of (1949) Anderson-Rubin statistic),

$$S(\beta) = T g_T(\beta)' \hat{\Sigma}_g(\beta)^{-1} g_T(\beta) \tag{3}$$

where $g_T(\beta) = \frac{1}{T} \sum Z_t (Y_t - \beta X_t)$ and $\hat{\Sigma}_g(\beta)$ is the usual heteroskedasticity-robust variance

[1] Since we consider the case with a single endogenous regressor, these robust F-statistic all coincide with the Kleibergen-Paap (2006) Wald statistic for testing that $\pi = 0$.

estimator for $\sqrt{T}g_T(\beta)$,

$$\hat{\Sigma}_g(\beta) = \frac{1}{T}\sum_t (Y_t - X_t\beta)^2 Z_t Z_t - g_T(\beta) g_T(\beta)'. \tag{4}$$

The S statistic evaluated at the true parameter value 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

$$CS_S = \left\{\beta : S(\beta) < \chi^2_{k,1-\alpha}\right\}$$

for $\chi^2_{k,1-\alpha}$ the $1-\alpha$ quantile of a χ^2_k distribution. Thus, the two-step confidence set based on the robust F-statistic F, cutoff c, Wald confidence set CS_W, and S confidence set CS_S, will be

$$CS_{2S} = \begin{cases} CS_W & \text{if } F \geq c \\ CS_S & \text{if } F < c \end{cases}. \tag{5}$$

In the remainder of this section, we examine the coverage of these two-step confidence sets, $Pr\{\beta_0 \in CS_{2S}\}$, for different Wald confidence sets and cutoffs. We first highlight that the conventional rule of thumb cutoff does not control coverage for two-step confidence sets even in homoskedastic models, while the SY cutoffs do. We then provide what is, to our knowledge, the first demonstration in the literature that the SY cutoffs fail to control coverage distortions under heteroskedasticity, even when using the heteroskedasticity-robust F-statistic.

For our simulations we set $\beta_0 = 0$ and assume that Z_t is a collection of dummy variables for different values of a categorical instrument $\tilde{Z}_t \in \{1, ..., k\}$. We take $k \in \{5, 10, 20\}$ and for each k consider two calibrations: one with a moderate degree of endogeneity (denoted by M) and the other with a very high degree of endogeneity (denoted by H). Our main simulation designs feature a substantial amount of heteroskedasticity, but we first examine the performance of two-step procedures under homoskedasticity. In each simulation calibration we consider a wide range of values for identification strength (as measured $\|\pi\|$) ranging from non-identification to very strong identification, and report the smallest coverage probability for each confidence set over these different values, $\min_{\|\pi\|} Pr_{\|\pi\|}\{\beta_0 \in CS\}$. All simulations

are based on samples of 10,000 observations. For further details on our simulation design see the Appendix.

2.1 The First Stage F-Statistic Under Homoskedasticity

We study the performance of nominal level 95% F-statistic-based two-step confidence sets (5) using different cutoffs c and different Wald confidence sets CS_W under homoskedasticity. We begin with the usual LIML and 2SLS confidence sets, which we can view as two-step confidence sets with $c = 0$. Next, we consider rule of thumb confidence sets which take $c = 10$. Finally we 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 when we take CS_W to be the 2SLS confidence set, and $c = 5.44$, $c = 3.68$, and $c = 3.21$ when we take CS_W to be the LIML confidence set.[2] 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%.

The results of this exercise are reported in Table 1. As these make clear, the rule of thumb cutoff of 10 does not ensure any fixed level of coverage control for two-step confidence sets: while 2SLS confidence sets based on the rule of thumb have coverage distortions less than 10% in the M calibrations, they exhibit more substantial distortions in the H calibrations, and the degree of distortion is increasing in the number of instruments k. In contrast, two-step confidence sets based on the cutoffs of SY have coverage distortions not exceeding 10% (and thus coverage not less than 85%) in all cases, as expected.

2.2 The First Stage F-Statistic Under Heteroskedasticity

As already noted, coverage control for two-step procedures based on the first stage F-statistic with cutoffs from SY relies critically on the assumption of homoskedasticity. To illustrate this fact we repeat the same simulation exercise as above but now take the errors to be heteroskedastic, so $Var\left((V_{1,t}, V_{2,t})\,|Z_t\right)$ depends on Z_t (see appendix for details). Since 2SLS and LIML are inefficient under heteroskedasticity, in addition to Wald confidence sets based

[2]We 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%.

	Medium Endogeneity (M)			High Endogeneity (H)		
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 1: Minimal coverage for nominal level 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 (5) 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, the robust S (or Anderson-Rubin) confidence set CS_S, and have asymptotic coverage at least 85% in models with homoskedastic errors.

on these estimators we also consider confidence sets based on CUGMM and 2SGMM. When considering two-step confidence sets based on SY we use LIML cutoffs for CUGMM and 2SLS cutoffs for 2SGMM.[3]

The minimal coverage for all confidence sets considered is reported in Table 2. 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, however, under heteroskedasticity two-step confidence sets using the SY cutoffs also fail to control coverage distortions, regardless of whether we use efficient or inefficient estimators. More generally, we see that in many cases heteroskedasticity gives rise to far more pronounced coverage shortfalls than we observed under homoskedasticity.

The central problem with two-step confidence sets based on the SY cutoffs is that under heteroskedasticity the first stage F-statistic is no longer a reliable indicator of identification strength, at least when used with conventional cutoffs. While this point has previously been highlighted with regard to 2SLS bias by Bun and de Haan (2010) and Olea and Pflueger (2013), the issue appears especially stark when considering coverage. In Figure 1 we plot the coverage of Wald confidence sets against the mean of the first-stage F-statistic for the model with ten instruments and the medium endogeneity (M) calibration as we vary $\|\pi\|$, noting that $E[F]$ is a strictly increasing function of $\|\pi\|$. As this figure makes clear, even when the

[3]We also considered CUGMM and 2SGMM in the homoskedastic case but, unsurprisingly given the large sample size, their behavior was indistinguishable from that of LIML and 2SLS respectively.

	Medium Endogeneity (M)			High Endogeneity (H)		
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 2: Minimal coverage for nominal level 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 (5) 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, the robust S confidence set CS_S, and have asymptotic coverage at least 85% in models with homoskedastic errors.

mean of the first stage F-statistic is 500, most nominal 95% Wald confidence sets exhibit coverage distortions exceeding 15%. A still more extreme version of this issue arises in the H calibration, where the 2SLS confidence set has a 15% coverage distortion even when the mean of the first stage F-statistic is 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.

3 A Simple Two-Step Confidence Set: A User's Guide

In this section we develop a simple approach to constructing two-step confidence sets which control coverage in general nonlinear GMM models. Our technique is based on the observation that one can construct a preliminary weak identification-robust confidence set $CS_{R,P}$ which is a strict subset of the conventional non-robust confidence sets CS_{NR} with probability tending to one under the usual strong identification asymptotics. Consequently, if we take ϕ_{ICS} to be an indicator for the event that $CS_{R,P} \not\subseteq CS_{NR}$ we have that $\phi_{ICS} \to_p 0$ under

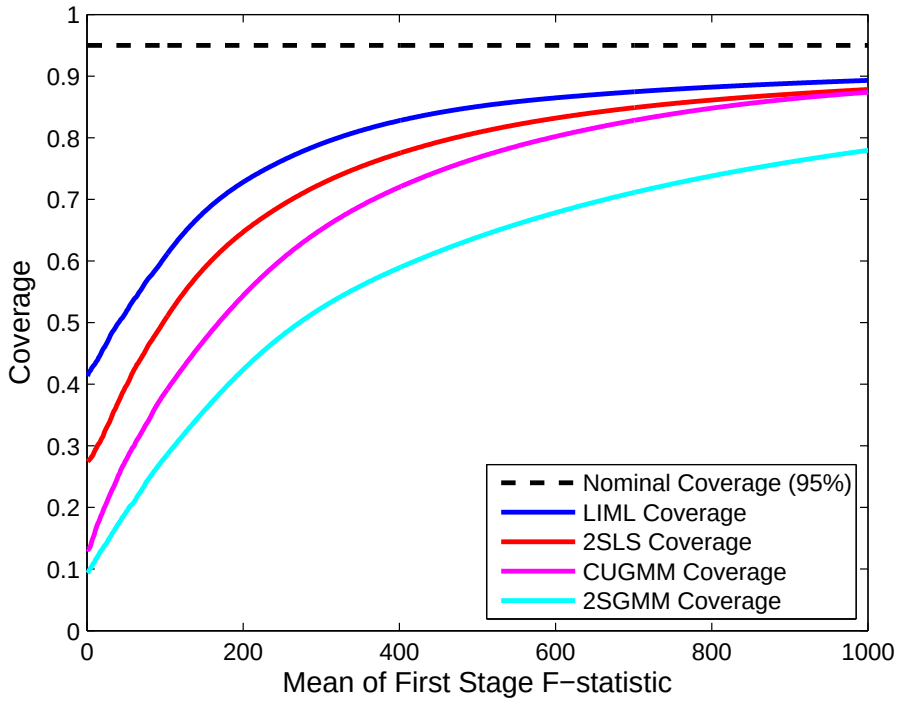

Figure 1: Coverage of Wald confidence sets plotted against the mean of the first stage F-statistic as we vary $\|\pi\|$ in heteroskedastic linear IV M calibration with $k = 10$.

strong identification. On the other hand provided our robust confidence set CS_R contains the preliminary confidence set, $CS_{R,P} \subseteq CS_R$, we can see that CS_{2S} will always contain the preliminary confidence set and thus will have coverage at least equal to that of $CS_{R,P}$. Hence, this approach yields two-step confidence sets which are equivalent to the usual, non-robust confidence set under strong identification while also controlling the maximal coverage distortion under weak identification.

This section focuses on the definition and implementation of our robust procedures, as well as some intuition for their properties. Formal discussion of the asymptotic properties of these procedures under weak- and strong-identification asymptotics are deferred to Sections 5 and 6 below.

3.1 GMM Model and Test Statistics

We consider a general GMM model with $k \times 1$-dimensional moment condition $g_t(\theta)$ which we assume is mean zero when the m-dimensional parameter θ is equal to its true value θ_0. In the

linear IV model, for example, we can take $\theta = \beta$ and $g_t(\beta) = Z_t(Y_t - X_t\beta)$. More generally, suppose we are interested in inference on a p-dimensional parameter ($p \leq m$) $\beta = f(\theta)$ for f a continuously differentiable function such that $\frac{\partial}{\partial \theta'} f(\theta_0)$ is full-rank. For example, we may be interested in constructing a confidence set for the ith element of the structural parameter vector and so take $f(\theta) = \theta_i$.

Let $g_T(\theta) = \frac{1}{T}\sum_t g_t(\theta)$ be the sample average of $g_t(\theta)$, and let $\hat{\Sigma}_g$, $\hat{\Sigma}_{\theta g}$, and $\hat{\Sigma}_\theta$ be estimators for $Var\left(\sqrt{T}g_T(\theta)\right)$, $Cov\left(\sqrt{T}vec\left(\frac{\partial}{\partial \theta'}g_T(\theta)\right), \sqrt{T}g_T(\theta)\right)$, and $Var\left(\sqrt{T}vec\left(\frac{\partial}{\partial \theta'}g_T(\theta)\right)\right)$ respectively, where $vec(A)$ denotes the vectorization of A, achieved by stacking its columns. In models where the data is independent across observations t we can use the usual heteroskedasticity-robust covariance estimators

$$\begin{aligned}
\hat{\Sigma}_g(\theta) &= \tfrac{1}{T}\textstyle\sum_t g_t(\theta) g_t(\theta)' - g_T(\theta) g_T(\theta)' \\
\hat{\Sigma}_{\theta g}(\theta) &= \tfrac{1}{T}\textstyle\sum_t \mathrm{vec}\left(\frac{\partial}{\partial \theta'} g_t(\theta)\right) g_t(\theta)' - \mathrm{vec}\left(\frac{\partial}{\partial \theta'} g_T(\theta)\right) g_T(\beta)' \\
\hat{\Sigma}_\theta(\theta) &= \tfrac{1}{T}\textstyle\sum_t \mathrm{vec}\left(\frac{\partial}{\partial \theta'} g_t(\theta)\right) \mathrm{vec}\left(\frac{\partial}{\partial \theta'} g_t(\theta)\right)' - \mathrm{vec}\left(\frac{\partial}{\partial \theta'} g_T(\theta)\right) \mathrm{vec}\left(\frac{\partial}{\partial \theta'} g_T(\theta)\right)'.
\end{aligned} \tag{6}$$

Define the S statistic as in (3), noting that (4) is simply $\hat{\Sigma}_g$ as defined in (6) specialized to the linear IV model. Following Kleibergen (2005), define a modified estimator for the mean of $\frac{\partial}{\partial \theta'} g_T(\theta)$,

$$D_T(\theta) = \left[\frac{\partial}{\partial \theta_1} g_T(\theta) - \hat{\Sigma}_{\theta_1 g}(\theta)\hat{\Sigma}_g(\theta)^{-1} g_T(\theta),\ \ldots,\ \frac{\partial}{\partial \theta_p} g_T(\theta) - \hat{\Sigma}_{\theta_p g}(\theta)\hat{\Sigma}_g(\theta)^{-1} g_T(\theta)\right]$$

where $\hat{\Sigma}_{\theta_i g}(\theta)$ is the $k \times k$ block of $\hat{\Sigma}_{\theta g}(\theta)$ corresponding to θ_i. Note that $vec(D_T(\theta)) = vec\left(\frac{\partial}{\partial \theta'} g_T(\theta)\right) - \hat{\Sigma}_{\theta g}(\theta)\hat{\Sigma}_g(\theta)^{-1} g_T(\theta)$. The K statistic of Kleibergen (2005) for testing hypotheses on the full parameter vector θ is then

$$K(\theta) = T g_T(\theta)' \hat{\Sigma}_g(\theta)^{-1} D_T(\theta) \left(D_T(\theta)' \hat{\Sigma}_g(\theta)^{-1} D_T(\theta)\right)^{-1} D_T(\theta)' \hat{\Sigma}_g(\theta)^{-1} g_T(\theta). \tag{7}$$

As discussed in Kleibergen (2005), $K(\theta)$ is a particular efficient GMM score statistic. Kleibergen shows that under mild regularity conditions $K(\theta_0)$ converges to a χ^2_m distribution in both weakly and strongly identified models, while the difference $S(\theta_0) - K(\theta_0)$ converges to a χ^2_{k-m} distribution and is asymptotically independent of $K(\theta_0)$.

We assume that we have some estimator $\tilde{\theta}$ for θ which under strong identification is first-order equivalent to

$$\hat{\theta} = \arg\min_{\theta} g_T(\theta)' \hat{\Omega}(\theta) g_T(\theta). \tag{8}$$

Here $\hat{\Omega}(\theta)$ is a symmetric positive-definite weighting matrix which we will assume converges uniformly in probability to a full-rank matrix $\Omega(\theta)$ under strong identification. Note that by taking $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$ this definition nests CUGMM as a special case. Since 2SGMM and CUGMM are first-order equivalent under strong identification this means we can also take $\tilde{\theta}$ to be the efficient two-step GMM estimator. Inefficient estimators are also allowed by (8), for instance the 2SLS estimator (which takes $\hat{\Omega}(\theta) = \left(\frac{1}{T}Z'Z\right)^{-1}$) applied to linear IV models with heteroskedastic errors.

The most common non-robust confidence set for β is based on the Wald statistic

$$W(\beta) = T\left(\tilde{\beta} - \beta\right)' \hat{\Sigma}_{\hat{\beta}}^{-1} \left(\tilde{\beta} - \beta\right) \tag{9}$$

where $\hat{\Sigma}_{\hat{\beta}}$ estimates the variance of $\sqrt{T}\tilde{\beta}$

$$\hat{\Sigma}_{\hat{\beta}} = \tfrac{\partial}{\partial\theta'} f\left(\tilde{\theta}\right)' \left(G_T\left(\tilde{\theta}\right)' \hat{\Omega}\left(\tilde{\theta}\right) G_T\left(\tilde{\theta}\right)\right)^{-1} G_T\left(\tilde{\theta}\right)' \hat{\Omega}\left(\tilde{\theta}\right) \hat{\Sigma}_g\left(\tilde{\theta}\right) \hat{\Omega}\left(\tilde{\theta}\right) G_T\left(\tilde{\theta}\right) \times \left(G_T\left(\tilde{\theta}\right)' \hat{\Omega}\left(\tilde{\theta}\right) G_T\left(\tilde{\theta}\right)\right)^{-1} \tfrac{\partial}{\partial\theta'} f\left(\tilde{\theta}\right)$$

for $G_T(\theta) = \frac{\partial}{\partial\theta'} g_T(\theta)$ the Jacobian of $g_T(\theta)$ with respect to θ. Under strong identification standard results yield that $W(\beta_0) \to_d \chi^2_p$.

3.2 A Generalized Weak Identification-Robust Test Statistic

To construct two-step confidence sets it will be useful to have a weak identification-robust test statistic which is asymptotically equivalent to the Wald statistic (9) under strong identification. The K statistic (7) satisfies this requirement if we are interested in inference on the full parameter vector $f(\theta) = \theta$ and use an efficient estimator with $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$, but since researchers frequently use inefficient estimators (e.g. 2SLS in linear IV with heteroskedasticity) and/or would like to conduct inference on some lower-dimensional function of the parameter vector, a more general robust statistic is needed. In this section we give a

heuristic derivation of a more general robust statistic whose properties are formally developed in Section 6.

A major difficulty with the Wald statistic (9) is that when identification is weak the estimator $\tilde{\theta}$ will typically be inconsistent with a highly non-standard distribution. Thus, rather than basing inference on the estimator $\tilde{\theta}$, robust statistics like (7) instead rely on the properties of the moment condition and its derivatives evaluated at a given parameter value θ, which can easily be studied even when identification is weak. We would like to follow this route and construct a statistic that depends only on the behavior of $g_T(\theta)$ and its derivative at θ, but that is also equivalent to $W(\beta)$ when identification is strong.

Happily for our purposes, under strong identification well known results (see e.g. Newey and McFadden 1994, Section 3.4) establish that the estimator $\tilde{\theta}$ is asymptotically equivalent to the one-step estimator based on any value θ in a $\sqrt{T}$ neighborhood of θ_0. Specifically, note that the first order condition for (8) implies that

$$\frac{\partial}{\partial\theta}\left(g_T\left(\hat{\theta}\right)'\hat{\Omega}\left(\hat{\theta}\right)g_T\left(\hat{\theta}\right)\right)=0$$

which, taking a mean-value expansion and discarding lower-order terms, yields that

$$G_T\left(\hat{\theta}\right)'\hat{\Omega}\left(\hat{\theta}\right)\left(g_T(\theta)+G_T(\theta^*)\left(\hat{\theta}-\theta\right)\right)+o_p\left(\frac{1}{\sqrt{T}}\right)=0$$

where θ^* is a value between $\hat{\theta}$ and θ that may vary across rows of $G_T(\theta^*)$ and $o_p\left(\frac{1}{\sqrt{T}}\right)$ denotes terms that converge in probability to zero at a rate faster than $\sqrt{T}$ under strong identification. Rearranging and imposing standard strong-identification assumptions (see Assumptions 5 and 6 below) this gives us that for θ in a $\sqrt{T}$-neighborhood of θ_0,

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

where the leading term on the right hand side is referred to as a one-step estimator. Note that this term depends only on the properties of the moment condition at parameter value θ, and so avoids issues arising from non-standard behavior for the estimator $\tilde{\theta}$. As noted in Kleibergen (2005), however, under weak identification $G_T(\theta)$ will in general be noisy and

correlated with $g_T(\theta)$. This can give rise to intractable behavior for the one-step estimator, but we can avoid these issues by replacing $G_T(\theta)$ with $D_T(\theta)$, which is asymptotically equivalent to $G_T(\theta)$ under strong identification but is asymptotically independent of $g_T(\theta)$. Using the Δ-method approximation $f\left(\hat{\theta}\right) - f(\theta) = \frac{\partial}{\partial \theta'} f(\theta)\left(\hat{\theta} - \theta\right) + o_p\left(\frac{1}{\sqrt{T}}\right)$ we then obtain a modified one-step estimator for $f\left(\hat{\theta}\right) - f(\theta)$:

$$f\left(\hat{\theta}\right) - f(\theta) = -\frac{\partial}{\partial \theta'} f(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} D_T(\theta)'\hat{\Omega}(\theta) g_T(\theta) + o_p\left(\frac{1}{\sqrt{T}}\right). \quad (10)$$

To construct our robust test statistic, we simply substitute the one step estimator in (10) into the expression for the Wald statistic (9), evaluate the covariance estimator $\hat{\Sigma}_{\hat{\beta}}$ at θ rather than $\hat{\theta}$, and replace $G_T(\theta)$ with $D_T(\theta)$ in all expressions. This yields the generalized robust statistic

$$\begin{array}{c} K_{\Omega,f}(\theta) = \\ Tg_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} \frac{\partial}{\partial \theta'} f(\theta)' \times \\ \left(\frac{\partial}{\partial \theta'} f(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} D_T(\theta)'\hat{\Omega}(\theta)\hat{\Sigma}_g(\theta) \times \right. \\ \left. \hat{\Omega}(\theta) D_T(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} \frac{\partial}{\partial \theta'} f(\theta)'\right)^{-1} \times \\ \frac{\partial}{\partial \theta'} f(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} D_T(\theta)'\hat{\Omega}(\theta) g_T(\theta). \end{array} \quad (11)$$

As our heuristic derivation highlights, we can think of $K_{\Omega,f}(\theta)$ as a Wald statistic based on a particular one-step estimator. Since this is one interpretation for $C(\alpha)$ statistics, $K_{\Omega,f}(\theta)$ is closely related to the GMM $C(\alpha)$ statistics discussed by Lee (2005). If we take $\hat{\Omega}(\theta)$ to be the efficient GMM weighting matrix, $K_{\Omega,f}(\theta)$ reduces to the $K(\theta)$ statistic (7) when we test the full parameter vector, while for $f(\theta)$ which selects a subvector of θ (e.g. the first parameter alone) one can show that $K_{\Omega,f}$ coincides with the LM_{eff} statistic proposed by Chaudhuri and Zivot (2011). We show in Section 6 that when evaluated at the true parameter value θ_0, $K_{\Omega,f}(\theta_0)$ converges to a χ^2_p distribution regardless of identification strength while $S(\theta_0) - K_{\Omega,f}(\theta_0)$ converges to a χ^2_{k-p} distribution and is asymptotically independent of $K_{\Omega,f}(\theta_0)$. Critically, under strong identification this statistic will be asymptotically equivalent to $W(f(\theta))$ on $\sqrt{T}$ neighborhoods of the true parameter value θ_0.

3.3 Our Suggested Two-Step Confidence Set

Using S and $K_{\Omega,f}$ we can now define our two-step confidence sets. For $a > 0$ define the preliminary robust confidence set as the set of values β such that there exists a value θ with $f(\theta) = \beta$ and $K_{\Omega,f}(\theta) + a \cdot S(\theta) \leq \chi^2_{p,1-\alpha}$,

$$CS_{R,P} = \left\{ \beta : \min_{\theta:\beta=f(\theta)} \left(K_{\Omega,f}(\theta) + a \cdot S(\theta) \right) \leq \chi^2_{p,1-\alpha} \right\}. \tag{12}$$

For an estimator $\tilde{\beta}$ which is first-order equivalent to (8) under strong identification, let $W(\beta)$ be the Wald statistic (9). Define CS_{NR} to be the Wald confidence set

$$CS_{NR} = \left\{ \beta : W(\beta) \leq \chi^2_{p,1-\alpha} \right\}, \tag{13}$$

and note that for $\tilde{\beta}$ 1-dimensional this gives us t-statistic confidence sets of the form (2). Let ϕ_{ICS} be an indicator function for $CS_{R,P} \not\subseteq CS_{NR}$

$$\phi_{ICS} = 1\{CS_{R,P} \not\subseteq CS_{NR}\}. \tag{14}$$

Define $F(x; a, k, p)$ to be the cumulative distribution function for a $(1+a)\chi^2_p + a\chi^2_{k-p}$ distribution and $F^{-1}(1-\alpha; a, k, p)$ to be the $1-\alpha$ quantile of this distribution. Our level $1-\alpha$ robust confidence set is

$$CS_R = \left\{ \beta : \min_{\theta:\beta=f(\theta)} \left(K_{\Omega,f}(\theta) + a \cdot S(\theta) \right) \leq F^{-1}(1-\alpha; a, k, p) \right\}. \tag{15}$$

We obtain the following as an immediate corollary of Theorems 1-3 below:

Corollary 1 *For $CS_{R,P}$, CS_{NR}, ϕ_{ICS}, and CS_R as defined in (12)-(15), the two step confidence set CS_{2S} defined in (1) has the following properties:*

1. *Under Assumptions 2-4 below, for $\gamma = 1 - \alpha - F\left(\chi^2_{p,1-\alpha}; a, k, p\right)$,*

$$\liminf_{T\to\infty} Pr\{\beta_0 \in CS_{2S}\} \geq F\left(\chi^2_{p,1-\alpha}; a, k, p\right) = 1 - \alpha - \gamma$$

under weak identification.

2. *Under strong identification (and Assumptions 2, 5, and 6 below),* $\phi_{ICS} \to_p 0$, $Pr\{CS_{2S} = CS_{NR}\} \to 1$, *and* $Pr\{\beta_0 \in CS_{2S}\} \to_p 1 - \alpha$.

Corollary 1 establishes that the recommended two-step confidence set has asymptotic coverage at least $1 - \alpha - \gamma$ regardless of identification strength. Moreover, CS_{2S} coincides with the the non-robust confidence set CS_{NR} with probability tending to one under strong identification, and thus has asymptotic coverage $1 - \alpha$ in this case. Thus, unlike two-step procedures based on the first-stage F-statistic, our approach controls coverage distortions in general models, including linear IV with heteroskedasticity.

3.4 Choosing the Value a and Computing Critical Values

The parameter $\gamma = 1 - \alpha - F\left(\chi^2_{p,1-\alpha}; a, k, p\right)$ measures the maximal coverage distortion for CS_{2S}. This is the price paid for using this two-step procedure, rather than using a fully robust confidence set like CS_R. Choosing a smaller value of γ reduces the maximal possible coverage distortion, but also makes our preliminary confidence set $CS_{R,P}$ larger, which in turn increases $Pr\{\phi_{ICS} = 1\}$ for any finite sample size and thus makes our ICS procedure more stringent, in the sense that it requires more extensive evidence before concluding that identification is strong.

Since $F\left(\chi^2_{p,1-\alpha}; a, k, p\right)$ is decreasing and continuous in a, converges to $1 - \alpha$ as $a \to 0$, and converges to 0 as $a \to \infty$, we can set a to achieve any value for γ between 0 and $1 - \alpha$. In practice, we suggest choosing γ and then selecting a accordingly. In particular, given k and p we recommend the following:

1. Select the coverage α for the non-robust confidence set and the maximal permitted coverage shortfall γ for the two-step procedure.

2. For M a large number, for each $m \in \{1, ..., M\}$ draw $(A_m, B_m) \sim \left(\chi^2_p, \chi^2_{k-p}\right)$ independently. Let $\tilde{a}(\gamma)$ solve

$$\frac{1}{M}\sum_{m=1}^{M} 1\left\{\left(1 + \tilde{a}(\gamma)\right) A_m + \tilde{a}(\gamma) B_m\right\} = 1 - \alpha - \gamma.$$

3. Let $\tilde{F}^{-1}\left(1 - \alpha; \tilde{a}(\gamma), k, p\right)$ equal the $1 - \alpha$ quantile of $\left(1 - \tilde{a}(\gamma)\right) A_m + \tilde{a}(\gamma) B_m$.

4. Construct the two-step confidence set CS_{2S} as above, using $a = \tilde{a}(\gamma)$ and approximating $F^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ by $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$.

As is easy to see, for M large this procedure will deliver two-step confidence sets with asymptotic coverage at least $1-\alpha-\gamma$ regardless of identification strength, and asymptotic coverage exactly $1-\alpha$ under strong identification. This procedure is easy to implement and the only user-selected parameter is the permitted level of coverage distortion γ. For ease of use, values for $\left(\tilde{a}(\gamma), \tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)\right)$ for $p \in \{1,2,3\}$ and $\alpha \in \{1\%, 5\%, 10\%\}$ are reported in Supplementary Appendix A. To avoid the use of non-standard critical values, we could also use alternative robust confidence sets whose construction requires only χ^2 critical values. See Supplementary Appendix B for one such approach.

4 Two-Step Confidence Sets in Linear IV

In this section, we apply the two-step confidence sets developed in Section 3 to the linear IV model discussed in Section 2. We show in simulation that our ICS procedure ϕ_{ICS} has performance competitive with pretests based on the first stage F-statistic in models with homoskedastic data but also controls coverage distortions for CS_{2S} even under heteroskedasticity.

4.1 Two Step Confidence Sets Under Homoskedasticity

We return to the homoskedastic IV model considered in Section 2.1 and simulate the coverage of the robust confidence sets CS_R and $CS_{R,P}$, as well as two-step confidence sets constructed as suggested in Section 3. For consistency with the simulations in Section 2.1, in all cases we set $\alpha = 5\%$ and $\gamma = 10\%$. By construction the robust confidence set CS_R has asymptotic coverage 95% under both weak and strong identification, while $CS_{R,P}$ has asymptotic coverage 85% and the two step confidence sets have minimal asymptotic coverage at least 85%. The simulation performance of these confidence sets, reported in Table 3, supports our theoretical results, showing that the simulated coverage of CS_R and $CS_{R,P}$ is quite close to their theoretical coverage while the minimal coverage of the two-step confidence sets is in all

	Medium Endogeneity (M)			High Endogeneity (H)		
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%
$CS_{R,P}$	84.7%	85%	85.2%	85.4%	85.6%	84.3%
CS_{2S} LIML	92.6%	92.4%	90.4%	86%	87%	85%
CS_{2S} 2SLS	92.8%	92.8%	92.9%	86%	87%	85%

Table 3: Minimal coverage for confidence sets in homoskedastic IV simulations with 10,000 observations, based on 2,500 simulations. CS_R and $CS_{R,P}$ 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_{2S} LIML and CS_{2S} 2SLS are two-step confidence sets (1) based on LIML and 2SLS, calculated as described in Section 3 for $\alpha = 5\%$ and $\gamma = 10\%$.

cases at least 85%.

Since pretesting procedures based on the critical values of SY (discussed in Section 4) also guarantee coverage at least 85% under homoskedasticity, it is interesting to compare their behavior to that of our ICS statistic (14). In Figure 2 we plot the mean of our ICS statistics $E[\phi_{ICS}]$ together with $E[\phi_{ICS,SY}]$, the mean of the ICS statistics based on the first stage F-statistic with SY's critical values, against the mean of the first-stage F-statistic as we vary $\|\pi\|$ in the moderate endogeneity (M) calibration with $k = 10$. As we can see, our ICS procedure for LIML behaves quite similarly to that of SY, while our ICS procedure for 2SLS indicates strong identification with substantially higher probability than that of SY. Repeating this exercise for the other M calibrations (results not shown) we find similar results, while when we consider the high endogeneity H calibrations we find no general ordering between our ICS procedures and those of SY.

4.2 Two Step Confidence Sets Heteroskedasticity

In this section we simulate the performance of our robust and two-step confidence sets in the heteroskedastic linear IV calibrations studied in Section 2.2. In particular, we consider the robust confidence sets CS_R and $CS_{R,P}$ 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 constructed as discussed in Section 3. In all cases, for consistency with Section 2.2 we take $\alpha = 5\%$ and $\gamma = 10\%$. We can see that as in the homoskedastic case discussed above our robust confidence sets

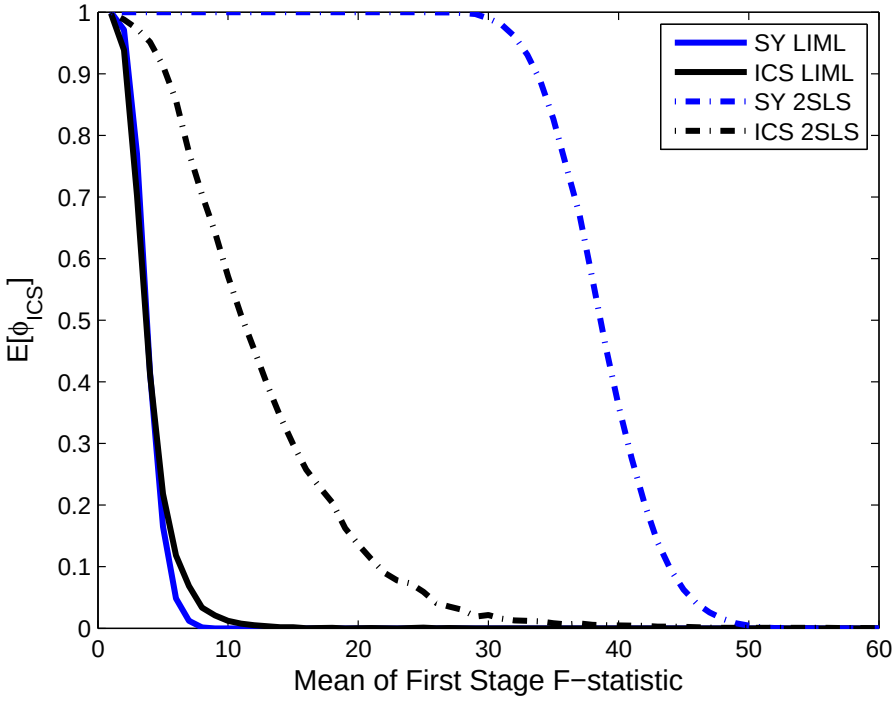

Figure 2: $E[\phi_{ICS}] = Pr\{\phi_{ICS} = 1\}$ plotted against the mean of the first stage F-statistic as we vary $\|\pi\|$ in heteroskedastic linear IV M calibration with $k = 10$, where SY LIML and SY 2SLS denote pretests based on the first stage F-statistic and the critical values of SY discussed in Section 2, while ICS LIML and ICS 2SLS use the ICS statistic (14) with $\alpha = 5\%$ and $\gamma = 10\%$.

	Medium Endogeneity (M)			High Endogeneity (H)		
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%
$CS_{R,P}$ Inefficient	84.7%	84.3%	85%	86%	85%	84.6%
CS_R Efficient	95.1%	94.6%	94.7%	94.9%	94.5%	95%
$CS_{R,P}$ Efficient	84.9%	84.5%	84.6%	84.8%	85.4%	84%
CS_{2S} LIML	94.1%	92.4%	93%	86.8%	86.8%	85.2%
CS_{2S} 2SLS	93.7%	94%	94.3%	86.7%	86.6%	85.2%
CS_{2S} CUGMM	95%	94.1%	93.4%	86.8%	92.6%	88.8%
CS_{2S} 2SGMM	94.5%	94%	93.9%	86.8%	92.8%	88.8%

Table 4: Minimal coverage for confidence sets in homoskedastic IV simulations with 10,000 observations, based on 2,500 simulations. CS_R and $CS_{R,P}$ Inefficient are robust 95% and 85% confidence sets (15) based on the two-stage least squares weight matrix $\hat{\Omega}(\theta) = \left(\frac{1}{T}Z'Z\right)^{-1}$, calculated as suggested in Section 3 for $\alpha = 5\%$ and $\gamma = 10\%$. CS_R and $CS_{R,P}$ Efficient are robust confidence sets with $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$ calculated as suggested in Section 3 $\alpha = 5\%$ and $\gamma = 10\%$. CS_{2S} LIML, CS_{2S} 2SLS, CS_{2S} CUGMM, and CS_{2S} 2SGMM are two-step confidence sets (1) based on LIML, 2SLS, CUGMM, and 2SGMM, calculated as described in Section 3 for $\alpha = 5\%$ and $\gamma = 10\%$.

CS_R and $CS_{R,P}$ have minimal coverage quite close to their theoretical coverage of 95% and 85%, respectively. Unlike the procedures based on the first stage F-statistic discussed in Section 2.2, we can also see that our two-step confidence sets have minimal coverage at least 85%, consistent with our theoretical results. Thus, our simulation results confirm that our approach delivers two-step confidence sets CS_{2S} which control coverage distortions, including in linear IV models with heteroskedasticity.

5 A Simple Two-Step Confidence Set: Theory

In this section we develop results concerning the properties of two-step confidence sets in general models. Our main results apply to contexts where we have a robust confidence set which is contained within the non-robust confidence set of interest with probability tending to one under strong identification. We show that whenever this condition holds we can construct a two-step confidence set which controls coverage distortions under weak identification and is asymptotically equivalent to the non-robust confidence set when identification is strong. Further, we note that these results can be used to create two-step confidence sets with uniformly bounded asymptotic coverage distortions when an appropriate uniformly correct confidence set is available. As a preliminary step we begin by defining several different

notions of limiting coverage probability, including asymptotic coverage and coverage under particular sequences of parameter values.

We assume that in sample T we observe data with distribution $F_T(\beta_{0,T}, \psi_{0,T})$, where $\beta \in B \subseteq \mathbb{R}^p$ is a parameter of interest and $\psi \in \Psi$ is some, potentially infinite-dimensional, nuisance parameter. We allow the true parameter values $\beta_{0,T}$ and $\psi_{0,T}$ to drift with the sample size, and index the sequence of true parameter values by

$$\xi = \{(\beta_{0,T}, \psi_{0,T})\}_{T=1}^{\infty} \in \Xi = \Pi_{T=1}^{\infty}(B \times \Psi).$$

We typically want confidence sets to control asymptotic coverage probability, defined as the limit of the minimal probability that CS contains the true value

$$ACP(CS) = \liminf_{T\to\infty} \inf_{(\beta_0,\psi_0)\in B\times\Psi} Pr_{T,(\beta_0,\psi_0)}\{\beta_0 \in CS\}.$$

In this paper, we will also be interested in less stringent notions of coverage. In particular, define the (asymptotic) sequential coverage probability of confidence set CS under the sequence of true parameter values ξ as

$$SCP(CS, \xi) = \liminf_{T\to\infty} Pr_{T,\xi}\{\beta_{0,T} \in CS\}.$$

Likewise, define the sequential coverage probability of confidence set CS under the set of sequences $\Xi^* \subset \Xi$ as the minimal sequential coverage probability under sequences in this set,

$$SCP(CS, \Xi^*) = \inf_{\xi\in\Xi^*} SCP(CS, \xi).$$

Note that sequential coverage probability under Ξ is simply the asymptotic coverage probability

$$SCP(CS, \Xi) = ACP(CS).$$

We will be particularly concerned with sequential coverage probability under "strongly identified sequences" which we will denote by $\Xi_S \subseteq \Xi$, and under "weakly identified sequences" which we will denote by $\Xi_W \subseteq \Xi$.

As above, suppose we have non-robust and robust confidence sets CS_{NR} and CS_R. We assume that our non-robust confidence set has sequential coverage at least $1-\alpha$ under strong identification

$$SCP\left(CS_{NR}, \Xi_S\right) \geq 1-\alpha, \tag{16}$$

but impose no restrictions on the performance of this confidence set under weak identification. By contrast, we will assume that the robust confidence set CS_R has coverage at least $1-\alpha$ under both weak and strong identification

$$\min\left\{SCP\left(CS_R, \Xi_S\right), SCP\left(CS_R, \Xi_W\right)\right\} \geq 1-\alpha. \tag{17}$$

For two-step confidence sets CS_{2S} as in (1), we can see that the sequential coverage of CS_{2S} will depend on the limiting behavior of ϕ_{ICS}. We can derive bounds which hold for all choices of ϕ_{ICS}, specifically

Lemma 1 *When (16) and (17) hold,*

1. $SCP\left(CS_{2S}, \Xi_W\right) \geq 1-\alpha-\sup_{\xi\in\Xi_W} \limsup_{T\to\infty} Pr_{T,\xi}\left\{\phi_{ICS}=0\right\}$
2. $SCP\left(CS_{2S}, \Xi_S\right) \geq 1-\alpha-\min\left\{\alpha, \sup_{\xi\in\Xi_S} \limsup_{T\to\infty} Pr_{T,\xi}\left\{\phi_{ICS}=1\right\}\right\}.$

These inequalities are tight, in the sense that one cannot obtain a sharper bound without additional restrictions on the joint behavior of CS_{NR}, CS_R, and ϕ_{ICS}.

Our approach to constructing two-step confidence sets exploits such additional structure. In particular, we make the following assumption:

Assumption 1 *We have a preliminary confidence set $CS_{R,P}$ such that:*

1. $SCP\left(CS_{R,P}, \Xi_W\right) \geq 1-\alpha-\gamma$
2. $Pr_{T,\xi}\left\{CS_{R,P} \subseteq CS_R\right\} = 1$ *for all* T *and* $\xi \in \Xi$
3. $\inf_{\xi\in\Xi_S} \liminf_{T\to\infty} Pr_{T,\xi}\left\{CS_{R,P} \subseteq CS_{NR}\right\} = 1.$

This assumption requires the existence of a preliminary confidence set that (1) has sequential coverage at least $1-\alpha-\gamma$ when identification is weak, (2) is contained in CS_R with probability

one, and (3) is contained in CS_{NR} with probability tending to one under all strongly identified sequences. While this might seem quite demanding, we show in the next section that for CS_{NR} a Wald confidence set and CS_R, $CS_{R,P}$ as defined in Section 3.3, this condition holds in general weakly identified GMM models under mild conditions. Under Assumption 1 we can easily establish that the two-step confidence sets CS_{2S} defined using ϕ_{ICS} as in (14) have bounded coverage distortions under weak identification and correct coverage under strong identification.

Theorem 1 *Under Assumption 1 together with (16), for ϕ_{ICS} as defined in (14) the two step confidence set CS_{2S} has the following properties:*

1. $SCP\left(CS_{2S}, \Xi_W\right) \geq 1 - \alpha - \gamma$
2. $SCP\left(CS_{2S}, \Xi_S\right) \geq 1 - \alpha$
3. $\inf_{\xi \in \Xi_S} \liminf_{T \to \infty} Pr_{T,\xi}\left\{CS_{2S} = CS_{NR}\right\} = 1.$

Further, $\sup_{\xi \in \Xi_S} \limsup_{T \to \infty} Pr_{T,\xi}\left\{\phi_{ICS} = 1\right\} = 0.$

Theorem 1 shows that whenever we have a preliminary robust confidence set satisfying Assumption 1 we can create a two-step confidence set which controls coverage distortions and is equal to our non-robust confidence set with probability tending to one under strong identification. This result highlights the value of identification-robust test statistics which are locally asymptotically equivalent to the conventional non-robust (e.g. Wald) statistics when identification is strong, since given such statistics it is typically straightforward to construct a preliminary confidence set satisfying Assumption 1 and thus to construct robust two-step confidence sets.

If there exists a preliminary confidence set $CS_{R,P}$ satisfying Assumption 1 which has uniformly correct coverage (one with $ACP\left(CS_{R,P}\right) \geq 1 - \alpha - \gamma$), by Theorem 1 we can use it to construct two-step confidence sets with uniformly correct coverage. Thus, since the results of D. Andrews, Cheng, and Guggenberger (2011) imply that $CS_{R,P}$ as defined in (12) has asymptotic coverage probability $1 - \alpha - \gamma$ in linear IV with one endogenous regressor

and heteroskedastic errors, Theorem 1 shows that the two-step confidence set developed in Section 3.3 has asymptotic coverage at least $1-\alpha-\gamma$ in this case, $ACP\left(CS_{2S}\right) \geq 1-\alpha-\gamma$.[4]

6 Two-Step Confidence Sets for GMM

Under mild assumptions, the preliminary robust confidence set $CS_{R,P}$ defined in (12) satisfies Assumption 1 in general GMM models, and thus can be used to construct two-step confidence sets which control coverage distortions. Specifically, as in Section 3, suppose we have a GMM model with moment condition $g_t\left(\theta\right)$ which is mean zero when evaluated at the true parameter value. To accommodate the drifting sequences of parameter values introduced in Section 5, let $E_{T,\xi}\left[\cdot\right]$ denote an expectation in sample size T under sequence of true parameter values ξ. We begin by giving conditions under which the robust confidence sets CS_R and $CS_{R,P}$ have correct sequential coverage even in weakly identified models, and then turn to establishing their properties in strongly identified models. Throughout this section, whenever we write a limit in an assumption we are implicitly assuming that limit exists.

6.1 Behavior Under Weak Identification

To establish the properties of our robust confidence sets under weak identification, we assume that a central limit theorem holds for the moment condition and its derivative evaluated at the true parameter value. In particular, much as in Kleibergen (2005) we assume that

Assumption 2 *For all $\xi \in \Xi_W \cup \Xi_S$, under ξ we have that for $J_{T,\xi}\left(\theta\right) = E_{T,\xi}\left[\frac{\partial}{\partial\theta'} g_T\left(\theta\right)\right]$,*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{pmatrix} g_t(\theta_{0,T}) \\ vec\left(\frac{\partial}{\partial\theta'} g_t(\theta_{0,T}) - J_{T,\xi}\left(\theta_{0,T}\right)\right)\end{pmatrix} \rightarrow_d \begin{pmatrix} \psi_g \\ \psi_{\theta_0}\end{pmatrix} \sim N\left(0, \begin{pmatrix} \Sigma_g & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_\theta\end{pmatrix}\right)$$

where Σ_g is positive definite and

$$\begin{pmatrix} \Sigma_g & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_\theta\end{pmatrix} = \lim_{T\rightarrow\infty} Var_{T,\xi}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{pmatrix} g_t(\theta_{0,T}) \\ vec\left(\frac{\partial}{\partial\theta'} g_t(\theta_{0,T})\right)\end{pmatrix}\right).$$

[4]To obtain this result, we impose the same restrictions on the nuisance parameter space Ψ as in D. Andrews, Cheng, and Guggenberger (2011).

Under this assumption we have that $\sqrt{T}g_T(\theta_{0,T}) = \frac{1}{\sqrt{T}}\sum_t g_t(\theta_{0,T})$ and its re-centered Jacobian converge jointly to a normal distribution. To construct test statistics with a known limiting distribution, we further need to assume that we can estimate the asymptotic variance matrix of $\left(\sqrt{T}g_T(\theta_{0,T}), \sqrt{T}vec\left(\frac{\partial}{\partial\theta'}g_T(\theta_{0,T})\right)\right)$ and that the weight matrix $\hat{\Omega}(\theta)$ converges to some non-random positive-definite limit.

Assumption 3 *For all $\xi \in \Xi_W$ we have estimators $\hat{\Sigma}_g(\theta_{0,T})$, $\hat{\Sigma}_{\theta g}(\theta_{0,T})$ and $\hat{\Sigma}_\theta(\theta_{0,T})$ which converge in probability to Σ_g, $\Sigma_{g\theta}$, and Σ_θ. Further, evaluated at the true parameter value the weight matrix $\hat{\Omega}(\theta)$ satisfies*

$$\hat{\Omega}(\theta_{0,T}) \to_p \Omega$$

for a non-stochastic symmetric positive-definite matrix Ω.

Provided Assumptions 2 and 3 hold, Lemma 1 of Kleibergen (2005) gives us a joint limiting distribution for the appropriately scaled and re-centered $g_T(\theta_{0,T})$ and $D_T(\theta_{0,T})$.

Lemma 2 *Under Assumptions 2 and 3 we have that for all $\xi \in \Xi_W$,*

$$\begin{pmatrix} \sqrt{T}g_T(\theta_{0,T}) \\ \sqrt{T}vec\left(D_T(\theta_{0,T}) - J_{T,\xi}(\theta_{0,T})\right) \end{pmatrix} \to_d \begin{pmatrix} \psi_g \\ \psi_D \end{pmatrix} \sim N\left(0, \begin{bmatrix} \Sigma_g & 0 \\ 0 & \Sigma_D \end{bmatrix}\right)$$

where $\Sigma_D = \Sigma_\theta - \Sigma_{\theta g}\Sigma_g^{-1}\Sigma_{g\theta}$.

To derive the limiting distribution for $K_{\Omega,f}$ as defined in (11), we need a final assumption to ensure that (i) we can normalize $D_T(\theta_{0,T})$ such that it converges to a limiting random variable which is non-singular almost surely and (ii) there exists a corresponding normalization for $\frac{\partial}{\partial\theta'}f(\theta_{0,T})$ which ensures that this term converges to a non-singular limit.

Assumption 4 *There exist sequences of full-rank normalizing matrices $\Lambda_{1,T}$ and $\Lambda_{2,T}$ of dimension $m \times m$ and $p \times p$, respectively, such that*

1. *$D_T(\theta)\Lambda_{1,T} \to_d D$ for a (possibly degenerate) Gaussian random matrix D which is full rank almost surely*

2. *$\Lambda_{2,T}\frac{\partial}{\partial\theta'}f(\theta_{0,T})\Lambda_{1,T} \to F$ for a full-rank matrix F*

Further, the elements of $\Lambda_{1,T}$ *are of order* $O\left(\sqrt{T}\right)$.[5]

This assumption is rather high-level, but can easily be verified in many leading cases. For example, Kleibergen (2005) considers the case where $\sqrt{T}J_{T,\xi}$ converges to a finite matrix J, in which case we can take $\Lambda_{1,T} = \sqrt{T}I_m$ and $\Lambda_{2,T} = \frac{1}{\sqrt{T}}I_p$. More broadly, this assumption holds under the commonly-used weakly identified GMM embedding of Stock and Wright (2000). Given these assumptions, it is easy to establish that $K_{\Omega,f}(\theta_0)$ and $S(\theta_0) - K_{\Omega,f}(\theta_0)$ have a well-behaved limiting distribution even under weak identification.

Theorem 2 *Under Assumptions 2, 3, and 4, under all* $\xi \in \Xi_W$,

$$\left(K_{\Omega,f}(\theta_{0,T}), S(\theta_{0,T}) - K_{\Omega,f}(\theta_{0,T})\right) \to_d \left(\chi^2_p, \chi^2_{k-p}\right)$$

and $K_{\Omega,f}(\theta_{0,T})$ *and* $S(\theta_{0,T}) - K_{\Omega,f}(\theta_{0,T})$ *are asymptotically independent.*

For tests of the full parameter vector ($f(\theta) = \theta$) based on the efficient weighting matrix $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$, Theorem 2 follows from results in Kleibergen (2005). Likewise, in the case where we are interested in testing a subset of the parameter vector (e.g. $f(\theta) = \theta_1$) and use the efficient weighting matrix $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$, Theorem 2 follows from results in Chaudhuri and Zivot (2011). The primary innovation of Theorem 2 is thus the fact that it allows for alternative weighting matrices $\hat{\Omega}(\theta) \neq \hat{\Sigma}_g(\theta)^{-1}$. While these weightings lead to inefficient tests and confidence sets when identification is strong, empirical researchers frequently use estimators based on inefficient weights (e.g. two-stage least squares applied to heteroskedastic data), so to construct two-step confidence sets based on such estimators it is important that we allow for general $\hat{\Omega}(\theta)$.

Note that Theorem 2 implies that $SCP(CS_{R,P}, \Xi_W) = 1-\alpha-\gamma$ and $SCP(CS_R, \Xi_W) = 1-\alpha$ for $CS_{R,P}$ and CS_R as defined in (12) and (15). Thus, since $CS_{R,P} \subseteq CS_R$ by definition, Assumption 1(1) and (2) hold. Hence, to verify Assumption 1 all that remains is to show that $CS_{R,P} \subseteq CS_{NR}$ with high probability under strong identification.

[5]That is, they are bounded above in absolute value by $C\sqrt{T}$ for some constant C.

6.2 Behavior Under Strong Identification

We now examine the properties of the preliminary confidence set $CS_{R,P}$ under strong identification. Specifically, we will define the set of strongly identified sequences Ξ_S as the set of sequences such that conventional GMM assumptions (closely related to those of e.g. Newey and McFadden (1994)) are satisfied. Following Newey and McFadden we consider two main groups of assumptions. The first group implies the consistency of the estimator

$$\hat{\theta} = \arg\min_{\theta} g_T(\theta)' \hat{\Omega}(\theta) g_T(\theta),$$

while the second group implies its asymptotic normality.[6] To simplify the exposition, we assume that for all $\xi \in \Xi_S$, $\theta_{0,T} = \theta_0$ for all T so that the true value θ_0 is not changing with the sample size.

Assumption 5 *For all $\xi \in \Xi_S$ the following conditions hold:*

1. $g_T(\theta) \to_p \lim_{T\to\infty} E_{T,\xi}[g_T(\theta)]$ *uniformly over the parameter space Θ for θ*
2. $E_{T,\xi}[g_T(\theta_0)] = 0 \ \forall T$
3. $\hat{\Omega}(\theta) \to_p \Omega(\theta)$ *uniformly over Θ for $\Omega(\theta)$ continuous and everywhere positive definite with a uniformly bounded maximal eigenvalue and a minimal eigenvalue bounded away from zero*
4. *For all $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\left(\lim_{T\to\infty} E_{T,\xi}[g_T(\theta)]\right)' \Omega(\theta) \left(\lim_{T\to\infty} E_{T,\xi}[g_T(\theta)]\right) < \delta$$

only if $\|\theta - \theta_0\| < \varepsilon$.

Assumption 5(1) requires that the sample mean of the moment condition $g_T(\theta_0)$ be uniformly close to its mean in large samples, while Assumption 5(3) requires that the weighting matrix be well-behaved. Assumption 5(2) and (4) are identification conditions, which ensure

[6] For simplicity we assume that we are working with CS_{NR} based directly on $\hat{\theta}$, but all results also apply to any $\tilde{\theta}$ which is first-order equivalent to $\tilde{\theta}$, that is $\tilde{\theta}$ such that $\|\hat{\theta} - \tilde{\theta}\| = o_p\left(\frac{1}{\sqrt{T}}\right)$ under strong identification.

that the population objective function is small if and only if evaluated in a neighborhood of the true parameter value. Assumption 5(4) will generally fail in contexts where identification issues arise. Provided these conditions hold, standard arguments yield the consistency of $\hat{\theta}$. Next, we consider an assumption yielding asymptotic normality of $\hat{\theta}$ with the usual limiting distribution.

Assumption 6 *The following conditions hold for all* $\xi \in \Xi_S$

1. θ_0 *belongs to the interior of* Θ
2. $g_T(\theta)$ *and* $\hat{\Omega}(\theta)$ *are almost surely continuously differentiable on some open ball* $B(\theta_0)$ *around* θ_0
3. *For*
$$J(\theta) = \lim_{T\to\infty} J_{T,\xi}(\theta) = \lim_{T\to\infty} E_{T,\xi}\left[\frac{\partial}{\partial\theta'} g_T(\theta)\right],$$
$J(\theta)$ *is continuous at* θ_0, $G_T(\theta) = \frac{\partial}{\partial\theta'} g_T(\theta) \to_p J(\theta)$ *uniformly on* $B(\theta_0)$, *and* $J(\theta_0)$ *is full-rank*
4. $\sup_{\theta\in B(\theta_0)} \left\| \frac{\partial vec\left(\hat{\Omega}(\theta)\right)}{\partial\theta'} \right\| = O_p(1)$
5. $\hat{\Sigma}_g(\theta) \to_p \Sigma_g(\theta)$ *uniformly on* $B(\theta_0)$, *and* $\Sigma_g(\theta) = \lim_{T\to\infty} Var_{T,\xi}(\sqrt{T} g_T(\theta))$ *is continuous in* θ *and everywhere positive-definite on* $B(\theta_0)$

Assumption 6(1) rules out cases where the true parameter value lies near the boundary of the parameter space. Assumption 6(2) requires that the moment condition and weight function both be smooth, while (3) and (4) require that their derivatives be well-behaved. Finally, Assumption 6(5) requires that we have a uniformly consistent estimator for $\Sigma_g(\theta)$ on a neighborhood of θ_0. Assumptions 2, 5, and 6 together yield the usual limiting distribution for the estimator $\hat{\theta}$ and the Wald statistic $W(\beta_0) = W(f(\theta_0))$. In particular, for $W(\beta)$ as in (9), standard arguments yield $W(f(\theta_0)) \to_d \chi^2_p$. Critically, these assumptions also imply the local asymptotic equivalence of the statistics $K_{\Omega,f}(\theta)$ and $W(f(\theta))$. Formally,

Lemma 3 *Let* $\{A_{\theta,T}\}$ *be a sequence of random sets such that* $\limsup_{T\to\infty} Pr\{A_{\theta,T} = \emptyset\} < 1$ *and* $\sup_{\theta\in A_{\theta,T}} \|\theta - \theta_0\| = O_p\left(\frac{1}{\sqrt{T}}\right)$ *(where we define the sup to be zero if* $A_{\theta,T}$ *is empty). Under Assumptions 2, 5, and 6, under all* $\xi \in \Xi_S$ *we have*

$$\sup_{\theta\in A_{\theta,T}} \|W(f(\theta)) - K_{\Omega,f}(\theta)\| = o_p(1).$$

Lemma 3 establishes that $K_{\Omega,f}(\theta)$ and $W(f(\theta))$ are locally asymptotically equivalent under strong identification, as suggested by our heuristic derivation in Section 3.2. Using this lemma, it is easy to verify Assumption 1(3).

Theorem 3 *Under Assumptions 2, 5, and 6, for* $CS_{R,P}$ *as defined in (12) and* CS_{NR} *as in (13) we have that*

$$\inf_{\xi\in\Xi_S} \liminf_{T\to\infty} Pr_{T,\xi}\{CS_{R,P} \subseteq CS_{NR}\} = 1.$$

Thus, under strong identification Assumption 1(3) holds. Since one can easily verify (16) under Assumptions 2-6, the conditions of Theorem 1 hold for the approach suggested in Section 3 in general GMM models.

7 Conclusion

In this paper we highlight that commonly-used approaches to detecting weak identification and constructing confidence sets can lead to highly misleading inferences. In particular we demonstrate that in linear IV two-step confidence sets based on the first-stage F-statistic can exhibit severe coverage distortions under heteroskedasticity. To remedy this issue we suggest an approach to detecting weak identification and constructing two-step confidence sets which we show controls coverage distortions in general GMM models. Applied to linear IV our approach yields two-step confidence sets that are competitive with those based on the first stage F-statistic under homoskedasticity, but which unlike conventional approaches also control coverage distortions under heteroskedasticity. While our results focus on the case where the non-robust component of our two-step confidence set is based on the Wald statistic, by the equivalence of the trinity of classical tests these results naturally extend to Quasi-Likelihood ratio or distance metric confidence sets.

While our simulations focus on linear IV, which is the most common context in which two-step procedures are currently applied, our theoretical results apply far more broadly and offer a novel approach to reliably assessing identification strength in a wide range of nonlinear and dynamic contexts where no alternative which controls coverage distortions for two-step confidence sets is currently available. Further, the core idea of our approach, that the equivalence of different confidence sets under the usual asymptotics allows us to asses the reliability of classical approximations, can also be applied to detecting the failure of standard confidence sets in other non-standard problems, for example robust inference in the presence of unit roots and inference on parameters near the boundary of the parameter space.

Appendix

IV Simulation Design

Heteroskedastic Case To examine the behavior of two-step confidence sets in simulation we need to specify the process generating (Z, V_1, V_2). Our focus is on heteroskedasticity, so we consider models where $(Z_t, V_{1,t}, V_{2,t})$ are independent across t but where $Var\left(V_{1,t}, V_{2,t} | \, Z_t\right)$ may depend on Z_t. We assume we have a categorical instrument and that Z_t is 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. We further assume that $\tilde{Z}_t$ is uniformly distributed so that $Pr\left\{Z_t = e_i\right\} = \frac{1}{k}$ for all $i \in \{1, ..., k\}$ and set the true parameter value $\beta_0 = 0$.

Since the support of Z_t is finite we can model $Var\left(V_{1,t}, V_{2,t} | \, Z_t\right)$ 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 we drew many values of $\Sigma_V\left(Z_t\right)$ and the direction of the first stage $\pi / \|\pi\|$ at random. For each draw we considered a large range of values for $||\pi||$, ranging from non-identification to very strong identification, and for our simulations we focus on particular draws of $\Sigma_V\left(Z_t\right)$ and $\pi / \|\pi\|$ that generate large coverage distortions for some values of $\|\pi\|$. We 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 $V_{2,t}$, which we denote by H, and the other with more moderate endogeneity, which we denote by M. The space of possible covariance structures is extremely large, however, so there certainly exist alternative parameter values generating much more pathological behavior for non-robust procedures than we report here. Consequently, our results give only lower bounds for possible coverage distortions. In all cases we consider simulated samples of 10,000 observations.

To give a sense of the parameter values used in our simulations, in Table 5 we report the

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

Table 5: Summary of linear IV calibration values. Note that $Corr(V_{1,t}, V_{2,t})$ is in all cases strictly less than one in absolute value, but the reported value is rounded to nearest 0.01.

(unconditional) correlation between $V_{1,t}$ and $V_{2,t}$ as well as $Stdev(Stdev(V_{i,t}|Z_t))/Stdev(V_{i,t})$, which is a natural measure for the degree of heteroskedasticity.

Homoskedastic Case

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

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

Proofs

Proof of Lemma 1

To prove (1), note that for any $\xi \in \Xi$ and any T, we have

$$Pr_{T,\xi}\{\beta_{0,T} \in CS_{2S}\} \geq Pr_{T,\xi}\{\beta_{0,T} \in CS_R\} - Pr_{T,\xi}\{\phi_{ICS} = 0\}.$$

By (17) $SCP(CS_R, \Xi_W) \geq 1-\alpha$, so Lemma 1(1) follows immediately from the definition of sequential coverage probability.

To prove (2), note that

$$Pr_{T,\xi}\{\beta_{0,T} \in CS_{2S}\} \geq Pr_{T,\xi}\{\beta_{0,T} \in CS_{NR}\} - Pr_{T,\xi}\{\beta_{0,T} \notin CS_R \text{ and } \phi_{ICS} = 1\},$$

and

$$Pr_{T,\xi}\{\beta_{0,T} \notin CS_R \text{ and } \phi_{ICS} = 1\} \leq \min\{Pr_{T,\xi}\{\beta_{0,T} \notin CS_R\}, Pr_{T,\xi}\{\phi_{ICS} = 1\}\}.$$

By (16) $SCP\left(CS_{NR},\Xi_S\right)\geq 1-\alpha$ so

$$SCP\left(CS_{2S},\Xi_S\right)\geq 1-\alpha-\sup_{\xi\in\Xi_S}\limsup_{T\to\infty}\min\left\{Pr_{T,\xi}\left\{\beta_{0,T}\notin CS_R\right\},Pr_{T,\xi}\left\{\phi_{ICS}=1\right\}\right\}$$

but $\sup_{\xi\in\Xi_S}\limsup_{T\to\infty}Pr_{T,\xi}\left\{\beta_{0,T}\notin CS_R\right\}\leq 1-\alpha$ by assumption, implying the result.

Proof of Theorem 1

To establish (1), note that by Assumption 1(2), $Pr_{T,\xi}\left\{CS_{R,P}\subseteq CS_R\right\}=1$ for all T and $\xi\in\Xi$. Thus by the definition of CS_{2S}, $Pr_{T,\xi}\left\{CS_{R,P}\subseteq CS_{2S}\right\}=1$ for all T and $\xi\in\Xi$. Consequently, $Pr_{T,\xi}\left\{\beta_{0,T}\in CS_{R,P}\right\}\leq Pr_{T,\xi}\left\{\beta_{0,T}\in CS_{2S}\right\}$, so (1) follows immediately from Assumption 1(1). (2) follows immediately from Lemma 1(2) and Assumption 1(3). (3) is implied by

$$\sup_{\xi\in\Xi_S}\limsup_{T\to\infty}Pr_{T,\xi}\left\{\phi_{ICS}=1\right\}=0,$$

which is an immediate consequence of Assumption 1(3).

Proof of Lemma 2 Follows immediately from Lemma 1 of Kleibergen (2005).

Proof of Theorem 2 Note that we can re-write $K_{\Omega,f}$ as

$$\begin{gathered}K_{\Omega,f}\left(\theta\right)=\\ Tg_T(\theta)'\hat{\Omega}\left(\theta\right)D_T(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}\left(\theta\right)D_T(\theta)\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta)'\Lambda_{2,T}'\times\\ \left(\Lambda_{2,T}\frac{\partial}{\partial\theta'}f(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}\left(\theta\right)D_T(\theta)\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}\left(\theta\right)\hat{\Sigma}(\theta)\times\right.\\ \left.\hat{\Omega}\left(\theta\right)D_T(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}\left(\theta\right)D_T(\theta)\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta)'\Lambda_{2,T}'\right)^{-1}\\ \Lambda_{2,T}\frac{\partial}{\partial\theta'}f(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}\left(\theta\right)D_T(\theta)\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'D_T\left(\theta\right)'\hat{\Omega}\left(\theta\right)g_T\left(\theta\right).\end{gathered}$$

By Lemma 2, $\left(\sqrt{T}g_T\left(\theta_{0,T}\right),\sqrt{T}\mathrm{vec}\left(D_T\left(\theta_{0,T}\right)-J_{T,\xi}\right)\right)$ converges to (ψ_g,ψ_D) which are mutually independent. Since we have assumed that the elements of $\Lambda_{1,T}$ are of order $\sqrt{T}$, note that $\frac{1}{\sqrt{T}}\Lambda_{1,T}=O\left(1\right)$, so $\left(\sqrt{T}g_T\left(\theta_{0,T}\right),D_T(\theta_{0,T})\Lambda_{1,T}\right)$ will be asymptotically independent as well. In particular, $\left(\sqrt{T}g_T\left(\theta_{0,T}\right),D_T(\theta_{0,T})\Lambda_{1,T}\right)\to_d\left(\psi_g,D\right)$ where $\psi_g|D\sim N\left(0,\Sigma_g\right)$.

We can further re-write $K_{\Omega,f}(\theta)$ as

$$Tg_T(\theta)'\hat{\Sigma}_g(\theta)^{-\frac{1}{2}} \times$$
$$P\left(\hat{\Sigma}_g(\theta)^{\frac{1}{2}}\hat{\Omega}(\theta)D_T(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta)D_T(\theta)\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta)'\Lambda_{2,T}'\right)\times$$
$$\hat{\Sigma}_g(\theta)^{-\frac{1}{2}}g_T(\theta).$$

where $P(X) = X(X'X)^{-1}X'$ denotes the projection matrix onto X. By Assumptions 3 and 4 and the Continuous Mapping Theorem,

$$\hat{\Sigma}_g(\theta_{0,T})^{\frac{1}{2}}\hat{\Omega}(\theta_{0,T})D_T(\theta_{0,T})\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta_{0,T})'\hat{\Omega}(\theta_{0,T})\hat{\Sigma}_g(\theta_{0,T})\hat{\Omega}(\theta_{0,T})D_T(\theta_{0,T})\Lambda_{1,T}\right)^{-1}\times$$
$$\Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta_{0,T})'\Lambda_{2,T} \to_d \Sigma_g^{\frac{1}{2}}\Omega D\left(D'\Omega\Sigma_g\Omega D\right)^{-1}F'$$

where the sole random component on the right hand side is D, and the right hand side has rank p almost surely. Together with the fact that $\Sigma_g^{-\frac{1}{2}}\psi_g|D \sim N(0, I_k)$, this implies by the Continuous Mapping Theorem that $\left(K_{\Omega,f}(\theta_{0,T}), D_T(\theta_{0,T})\Lambda_{1,T}\right) \to_d \left(K_{\Omega,f}^*, D\right)$ where $K_{\Omega,f}^*|D \sim \chi_p^2$, since conditional on D $K_{\Omega,f}^*$ is a quadratic form in a standard-normal random vector and a rank-p projection matrix.

We can handle $S(\theta_{0,T}) - K_{\Omega,f}(\theta_{0,T})$ in a similar manner. In particular, note that

$$S(\theta) - K_{\Omega,f}(\theta) =$$
$$Tg_T(\theta)'\hat{\Sigma}_g(\theta)^{-\frac{1}{2}} \times$$
$$\left(I - P\left(\hat{\Sigma}_g(\theta)^{\frac{1}{2}}\hat{\Omega}(\theta)D_T(\theta)\Lambda_{1,T}\left(\Lambda_{1,T}'D_T(\theta)'\hat{\Omega}(\theta)\hat{\Sigma}_g(\theta)\hat{\Omega}(\theta)D_T(\theta)\Lambda_{1,T}\right)^{-1}\Lambda_{1,T}'\frac{\partial}{\partial\theta'}f(\theta)'\Lambda_{2,T}'\right)\right)\times$$
$$\hat{\Sigma}_g(\theta)^{-\frac{1}{2}}g_T(\theta).$$

so

$$\left(K_{\Omega,f}(\theta_{0,T}), S(\theta_{0,T}) - K_{\Omega,f}(\theta_{0,T}), D_T(\theta_{0,T})\Lambda_{1,T}\right) \to_d \left(K_{\Omega,f}^*, S^* - K_{\Omega,f}^*, D\right)$$

where $\left(K_{\Omega,f}^*, S^* - K_{\Omega,f}^*\right)|D \sim \left(\chi_p^2, \chi_{k-p}^2\right)$ and $\left(K_{\Omega,f}^*, S^* - K_{\Omega,f}^*\right)$ are independent conditional on D. Thus we have that $\left(K_{\Omega,f}^*, S^* - K_{\Omega,f}^*\right)$ and independent and distributed $\left(\chi_p^2, \chi_{k-p}^2\right)$ unconditionally as well, which establishes the result.

Proof of Lemma 3 The proof is standard, but is included for completeness. If $A_{\theta,T}$ is empty, we have defined $\sup_{\theta\in A_{\theta,T}}\|W(f(\theta)) - K_{\Omega,f}(\theta)\| = 0$. Hence, we restrict atten-

tion 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, $Pr\{A_{\theta,T} \subset B(\theta_0)\} \to 1$. Hence, by our 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

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

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:

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

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,

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

Similar arguments establish that

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

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

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

under our assumptions, this suffices to establish the desired equivalence for tests of the full

parameter vector, which take $f(\theta) = \theta$. To complete the proof, we 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 our assumptions

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

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

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

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,

$$\sup_{\theta \in A_{\theta,T}} \left\| \left(\frac{\partial}{\partial \theta'} f(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} D_T(\theta)'\hat{\Omega}(\theta)\hat{\Sigma}_g(\theta)\hat{\Omega}(\theta)\hat{D}_T(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} \frac{\partial}{\partial \theta'} f(\theta)' \right)^{-1} - \left(\frac{\partial}{\partial \theta'} f(\theta_0)\left(J(\theta_0)'\Omega(\theta_0) J(\theta_0)\right)^{-1} J(\theta_0)'\hat{\Omega}(\theta_0)\hat{\Sigma}_g(\theta_0)^{-1}\hat{\Omega}(\theta_0) J(\theta_0)\left(J(\theta_0)'\Omega(\theta_0) J(\theta_0)\right)^{-1} \frac{\partial}{\partial \theta'} f(\theta_0)' \right)^{-1} \right\| = o_p(1)$$

by the Continuous Mapping Theorem, and by the triangle inequality

$$\sup_{\theta \in A_{\theta,T}} \left\| \left(\frac{\partial}{\partial \theta'} f(\hat{\theta})\left(G_T(\hat{\theta})'\hat{\Omega}\left(\hat{\theta}\right) G_T(\hat{\theta})\right)^{-1} G_T(\hat{\theta})'\hat{\Omega}\left(\hat{\theta}\right)' \hat{\Sigma}_g(\hat{\theta})^{-1}\hat{\Omega}\left(\hat{\theta}\right) G_T(\hat{\theta})\left(G_T(\hat{\theta})'\hat{\Omega}\left(\hat{\theta}\right) G_T(\hat{\theta})\right)^{-1} \frac{\partial}{\partial \theta'} f(\hat{\theta})' \right)^{-1} - \left(\frac{\partial}{\partial \theta'} f(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} D_T(\theta)'\hat{\Omega}(\theta)\hat{\Sigma}_g(\theta)^{-1}\hat{\Omega}(\theta) D_T(\theta)\left(D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta)\right)^{-1} \frac{\partial}{\partial \theta'} f(\theta)' \right)^{-1} \right\| = o_p(1).$$

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

Proof of Theorem 3 For $S(\theta)$ as in (3), note that Assumption 5 implies that for any $\varepsilon > 0$,

$$\inf_{\|\theta-\theta_0\|\geq\varepsilon} S(\theta) \to_p \infty.$$

Thus, if we define $A_{\theta,T} = \left\{\theta : a \cdot S(\theta) \leq \chi^2_{p,1-\alpha}\right\}$, we can see that $\sup_{\theta\in A_{\theta,T}} \|\theta - \theta_0\| = o_p(1)$. A mean-value expansion yields that $g_T(\theta) = g_T(\theta_0) + G_T(\theta^*)(\theta - \theta_0)$ for θ^* an intermediate value which may vary across rows. Since $\sup_{\theta\in B(\theta_0)} \|G_T(\theta) - J(\theta)\| = o_p(1)$, and $\sup_{\theta\in B(\theta_0)} \left\|\hat{\Sigma}_g(\theta) - \Sigma_g(\theta)\right\| = o_p(1)$ for an open ball $B(\theta_0)$ around θ_0 as in Assumption 6 and $J(\theta)$ and $\Sigma_g(\theta)$ are continuous in θ, we can see that

$$\sup_{\theta\in A_{\theta,T}} \left|S(\theta) - T\left(g_T(\theta_0) + J(\theta_0)(\theta - \theta_0)\right)' J(\theta_0)' \Sigma_g(\theta_0)^{-1} \left(g_T(\theta_0) + J(\theta_0)(\theta - \theta_0)\right)\right| = o_p(1).$$

Thus, for any $\varepsilon > 0$ we can see that for $\underline{\lambda}$ the minimal eigenvalue of $\Sigma_g(\theta_0)^{-1}$,

$$Pr_{T,\xi}\left\{\inf_{\theta\in A_{\theta,T}} \left(S(\theta) - \underline{\lambda} T \|g_T(\theta_0) + J(\theta_0)(\theta - \theta_0)\|^2\right) > -\varepsilon\right\} \to 1.$$

Since $\sqrt{T} g_T(\theta_0) = O_p(1)$ by Assumption 2, this implies thatsup$_{\theta\in A_{\theta,T}} \|\theta - \theta_0\| = o_p\left(\frac{1}{\sqrt{T}}\right)$. Thus, we have established that $A_{\theta,T} = \left\{\theta : a \cdot S(\theta) \leq \chi^2_{p,1-\alpha}\right\}$ shrinks towards θ_0 at rate $\sqrt{T}$.

Next, note that $K_{\Omega,f}(\theta) \geq 0$ by construction, so $K_{\Omega,f}(\theta) + a \cdot S(\theta) \geq a \cdot S(\theta)$ and $CS_{R,P} \subseteq A_{\theta,T}$. By standard results on the distribution of tests for over-identifying restrictions $\inf_\theta S(\theta) \to_d \chi^2_{k-p_\theta}$, so since

$$K_{\Omega,f}(\theta) + a \cdot S(\theta) \geq K_{\Omega,f}(\theta) + a \cdot \inf_\theta S(\theta)$$

and by Lemma 3 we know that $\sup_{\theta\in A_{\theta,T}} |K_{\Omega,f}(\theta) - W(f(\theta))| = o_p(1)$, we obtain

$$Pr_{T,\xi}\left\{\inf_{\theta\in A_{\theta,T}} \left(K_{\Omega,f}(\theta) + a \cdot S(\theta) - W(f(\theta))\right) > 0\right\} \to 1$$

with the consequence that $Pr_{T,\xi}\{CS_{R,P} \subseteq CS_{NR}\} \to 1$, as we wanted to show.

References

Anderson T.W. and H. Rubin (1949): "Estimators for the Parameters of a Single Equation in a Complete Set of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 570-582.

Andrews D.W.K, and X. Cheng (2012): "Estimation and Inference with Weak, Semi-strong, and Strong Identification," *Econometrica*, 80, 2153-2211.

Andrews D.W.K., X. Cheng, and P. Guggenberger (2011): "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," Cowles Foundation working paper.

Arellano M., L.P. Hansen and E. Sentana (2012): "Underidentification?" *Journal of Econometrics*, 170, 256-280.

Bun M. and M. de Haan (2010): "Weak Instruments and the first-stage F-statistic in IV Models with a Nonscalar Error Covariance Structure," Unpublished Manuscript.

Chaudhuri S. and E. Zivot (2011): "A New Method of Projection-Based Inference in GMM with Weakly Identified Nuisance Parameters", *Journal of Econometrics*, 164, 239-251.

Elliott G., U.K. Mueller, and M.W. Watson (2012): "Nearly Optimal Tests When a Nuisance Parameter is Present Under The Null Hypothesis," Unpublished Manuscript.

Hahn J. and J. Hausman (2002): "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, 70, 163-189.

Hausman J., J. Stock, and M. Yogo (2005): "Asymptotic Properties of the Hahn-Hausman Test for Weak Instruments," *Economics Letters*, 89, 333-342.

Inoue A. and B. Rossi (2011): "Testing for Weak Identification in Possibly Nonlinear Models," *Journal of Econometrics*, 161, 246-261.

Kleibergen F. (2005): "Testing Parameters in GMM Without Assuming They Are Identified," *Econometrica*, 73, 1103-1123.

Kleibergen F. and Paap R. (2006): "Generalized Reduced Rank Tests Using the Singular Value Decomposition," *Journal of Econometrics*, 133, 97-126.

Lee, L. (2005) "A $C(\alpha)$-type Gradient Test in the GMM Approach," unpublished manuscript.

Newey W. and D.L. McFadden (1994): "Large Sample Estimation and Hypothesis Test-

ing," Chapter 36 of D.L. McFadden and R.F. Engle Eds, *Handbook of Econometrics, Volume 4*, Elsever.

Olea J.L.M. and C. Pflueger (2013): "A Robust Test for Weak Instruments," *Journal of Business and Economic Statistics*, 31, 358-369.

Staiger D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557-586.

Stock J. and J. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055–1096.

Stock J., J. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Review of Economics and Statistics*, 20, 518-529.

Stock, J. and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg,* D. W. K. Andrews and J. Stock, eds., Cambridge: Cambridge University Press, pp. 80-108.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

Wright J. (2003): "Detecting Lack of Identification in GMM," *Econometric Theory*, 19, 322-330.

Wright J. (2010): "Testing the Adequacy of Conventional Asymptotics in GMM," *Econometrics Journal*, 13, 205-217.

Supplementary Appendix to

Robust Two-Step Confidence Sets, and the Trouble with the First Stage F-Statistic

Isaiah Andrews

Section A in this Supplementary Appendix tabulates weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha;\tilde{a}(\gamma),k,p)$ for calculating the robust confidence sets $CS_{R,P}$ and CS_R defined in (12) and (15) in the main text. Section B discusses an alternative approach to creating robust confidence sets which requires only χ^2 critical values.

Supplementary Appendix A: Critical Values

Tables 6-14 report weights $\tilde{a}(\gamma)$ for constructing robust confidence sets $CS_{R,P}$ and CS_R as in (12) and (15), together with critical values $\tilde{F}^{-1}(1-\alpha;\tilde{a}(\gamma),k,p)$ for calculating CS_R. These values are calculated following the approach discussed in Section 3.4 based on ten million simulation draws ($M = 10^7$). Each table corresponds to a given pair (α,p) for $\alpha \in \{0.01, 0.05, 0.1\}$ and $p \in \{1,2,3\}$ and reports values for number of moment conditions $k \in \{1,2,...,30\}$ and maximal distortion $\gamma \in \{0.01, 0.05, 0.1, 0.15, 0.2\}$.

Supplementary Appendix B: Alternative Robust Confidence Sets

The robust confidence sets $CS_{R,P}$ and CS_R defined in (12) and (15) represent only one of many ways to construct robust confidence sets which satisfy Assumption 1 and that can thus be used to construct two-step confidence sets. While we find that this approach performs quite well in simulation, it has the disadvantage of requiring that we calculate appropriate weights $a(\gamma)$ and critical values $F^{-1}(1-\alpha;\tilde{a}(\gamma),k,p)$ (though we tabulate these for many values of (α,γ,k,p) in Supplementary Appendix A). In this section, we discuss an alternative approach to constructing robust confidence sets, based on a generalization of the JK tests suggested by Kleibergen (2005), which requires only χ^2 critical values.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
1	0.221	8.15	0.871	12.48	1.593	17.30	2.354	22.38	3.216	28.13
2	0.185	8.10	0.623	11.63	0.986	14.72	1.311	17.55	1.633	20.40
3	0.159	8.08	0.498	11.24	0.744	13.68	0.947	15.78	1.138	17.79
4	0.140	8.06	0.419	10.99	0.605	13.09	0.751	14.80	0.885	16.41
5	0.124	8.03	0.363	10.82	0.514	12.70	0.628	14.19	0.729	15.53
6	0.113	8.01	0.321	10.67	0.449	12.42	0.541	13.74	0.623	14.93
7	0.102	7.99	0.289	10.58	0.399	12.21	0.477	13.44	0.544	14.50
8	0.094	7.99	0.263	10.52	0.360	12.08	0.428	13.20	0.484	14.17
9	0.087	7.98	0.241	10.45	0.328	11.94	0.387	12.99	0.437	13.90
10	0.081	7.97	0.223	10.40	0.302	11.83	0.355	12.83	0.399	13.68
11	0.076	7.96	0.207	10.34	0.280	11.73	0.327	12.69	0.367	13.50
12	0.071	7.95	0.193	10.30	0.261	11.66	0.304	12.58	0.339	13.34
13	0.067	7.95	0.182	10.26	0.244	11.59	0.284	12.47	0.316	13.21
14	0.064	7.95	0.171	10.23	0.230	11.54	0.267	12.40	0.296	13.10
15	0.060	7.94	0.162	10.20	0.217	11.49	0.251	12.32	0.278	13.00
16	0.057	7.94	0.154	10.18	0.205	11.45	0.238	12.26	0.263	12.92
17	0.055	7.94	0.146	10.15	0.195	11.40	0.225	12.19	0.249	12.83
18	0.053	7.94	0.140	10.13	0.186	11.37	0.214	12.15	0.236	12.76
19	0.050	7.93	0.134	10.12	0.178	11.34	0.205	12.11	0.225	12.71
20	0.048	7.93	0.128	10.10	0.170	11.30	0.196	12.06	0.215	12.64
21	0.046	7.93	0.123	10.10	0.163	11.28	0.187	12.03	0.206	12.60
22	0.045	7.92	0.118	10.08	0.157	11.26	0.180	11.99	0.197	12.55
23	0.043	7.93	0.114	10.07	0.151	11.25	0.173	11.96	0.189	12.52
24	0.041	7.93	0.110	10.06	0.145	11.24	0.166	11.94	0.182	12.48
25	0.040	7.93	0.106	10.05	0.140	11.22	0.161	11.92	0.176	12.45
26	0.039	7.93	0.102	10.04	0.135	11.21	0.155	11.90	0.170	12.42
27	0.038	7.93	0.099	10.04	0.131	11.19	0.150	11.88	0.164	12.40
28	0.036	7.92	0.096	10.03	0.127	11.17	0.145	11.86	0.158	12.37
29	0.035	7.92	0.093	10.02	0.123	11.15	0.141	11.84	0.153	12.34
30	0.034	7.92	0.090	10.01	0.119	11.15	0.136	11.82	0.149	12.32

Table 6: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ for $p = 1$ and $\alpha = 0.01$.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
1	0.085	4.17	0.419	5.45	0.852	7.12	1.338	8.99	1.901	11.15
2	0.066	4.17	0.289	5.30	0.510	6.48	0.710	7.60	0.911	8.75
3	0.054	4.16	0.226	5.23	0.378	6.24	0.504	7.12	0.623	7.97
4	0.046	4.16	0.186	5.18	0.304	6.11	0.396	6.86	0.479	7.56
5	0.040	4.16	0.159	5.15	0.256	6.02	0.329	6.70	0.392	7.32
6	0.035	4.16	0.138	5.13	0.221	5.96	0.282	6.59	0.333	7.15
7	0.031	4.16	0.123	5.11	0.196	5.92	0.247	6.52	0.290	7.03
8	0.028	4.15	0.111	5.10	0.175	5.88	0.221	6.46	0.258	6.94
9	0.026	4.15	0.101	5.09	0.159	5.86	0.199	6.41	0.232	6.87
10	0.024	4.15	0.092	5.08	0.146	5.84	0.182	6.37	0.211	6.82
11	0.022	4.15	0.085	5.07	0.134	5.82	0.167	6.34	0.194	6.77
12	0.021	4.15	0.079	5.06	0.125	5.80	0.155	6.32	0.179	6.73
13	0.019	4.15	0.074	5.06	0.116	5.79	0.145	6.30	0.167	6.70
14	0.018	4.15	0.069	5.05	0.109	5.78	0.136	6.28	0.156	6.67
15	0.017	4.15	0.065	5.05	0.103	5.77	0.127	6.26	0.146	6.65
16	0.016	4.15	0.062	5.04	0.097	5.76	0.120	6.25	0.138	6.63
17	0.015	4.15	0.059	5.04	0.092	5.75	0.114	6.24	0.131	6.61
18	0.015	4.15	0.056	5.04	0.087	5.75	0.108	6.23	0.124	6.59
19	0.014	4.15	0.053	5.04	0.083	5.74	0.103	6.22	0.118	6.58
20	0.013	4.15	0.051	5.03	0.079	5.74	0.098	6.21	0.113	6.57
21	0.013	4.15	0.048	5.03	0.076	5.73	0.094	6.20	0.108	6.56
22	0.012	4.15	0.046	5.03	0.073	5.73	0.090	6.20	0.103	6.55
23	0.012	4.15	0.045	5.03	0.070	5.72	0.087	6.19	0.099	6.54
24	0.011	4.15	0.043	5.02	0.067	5.72	0.083	6.18	0.095	6.53
25	0.011	4.15	0.041	5.02	0.065	5.71	0.080	6.18	0.092	6.52
26	0.011	4.15	0.040	5.02	0.063	5.71	0.078	6.18	0.088	6.52
27	0.010	4.15	0.039	5.02	0.060	5.71	0.075	6.17	0.085	6.51
28	0.010	4.15	0.037	5.02	0.058	5.70	0.072	6.17	0.083	6.51
29	0.010	4.15	0.036	5.02	0.057	5.70	0.070	6.16	0.080	6.50
30	0.009	4.15	0.035	5.02	0.055	5.70	0.068	6.16	0.077	6.50

Table 7: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ for $p = 1$ and $\alpha = 0.05$.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
1	0.058	2.87	0.305	3.53	0.647	4.46	1.043	5.53	1.517	6.82
2	0.042	2.86	0.196	3.46	0.359	4.13	0.510	4.79	0.665	5.47
3	0.033	2.86	0.147	3.43	0.258	4.02	0.352	4.55	0.442	5.07
4	0.027	2.86	0.118	3.41	0.204	3.96	0.273	4.43	0.335	4.87
5	0.023	2.86	0.099	3.40	0.170	3.93	0.224	4.36	0.272	4.75
6	0.020	2.86	0.085	3.39	0.145	3.90	0.191	4.31	0.229	4.67
7	0.017	2.86	0.075	3.38	0.127	3.89	0.166	4.27	0.199	4.61
8	0.016	2.86	0.067	3.38	0.114	3.87	0.148	4.25	0.176	4.57
9	0.014	2.86	0.061	3.38	0.102	3.86	0.133	4.23	0.158	4.54
10	0.013	2.86	0.055	3.37	0.093	3.86	0.121	4.22	0.143	4.52
11	0.012	2.86	0.051	3.37	0.086	3.85	0.111	4.21	0.131	4.50
12	0.011	2.86	0.047	3.37	0.079	3.84	0.103	4.20	0.121	4.48
13	0.010	2.86	0.044	3.37	0.074	3.84	0.095	4.19	0.112	4.47
14	0.010	2.86	0.041	3.36	0.069	3.83	0.089	4.18	0.105	4.46
15	0.009	2.86	0.038	3.36	0.065	3.83	0.084	4.18	0.098	4.45
16	0.008	2.86	0.036	3.36	0.061	3.83	0.079	4.17	0.093	4.44
17	0.008	2.86	0.034	3.36	0.058	3.82	0.075	4.17	0.088	4.43
18	0.008	2.86	0.032	3.36	0.055	3.82	0.071	4.16	0.083	4.43
19	0.007	2.86	0.031	3.36	0.052	3.82	0.067	4.16	0.079	4.42
20	0.007	2.86	0.029	3.36	0.050	3.82	0.064	4.16	0.075	4.42
21	0.007	2.86	0.028	3.36	0.047	3.82	0.061	4.15	0.072	4.41
22	0.006	2.86	0.027	3.36	0.045	3.81	0.059	4.15	0.069	4.41
23	0.006	2.86	0.026	3.36	0.043	3.81	0.056	4.15	0.066	4.41
24	0.006	2.86	0.025	3.35	0.042	3.81	0.054	4.15	0.063	4.40
25	0.006	2.86	0.024	3.35	0.040	3.81	0.052	4.14	0.061	4.40
26	0.005	2.86	0.023	3.35	0.039	3.81	0.050	4.14	0.059	4.40
27	0.005	2.86	0.022	3.35	0.037	3.81	0.048	4.14	0.057	4.40
28	0.005	2.86	0.021	3.35	0.036	3.80	0.047	4.14	0.055	4.39
29	0.005	2.86	0.021	3.35	0.035	3.80	0.045	4.14	0.053	4.39
30	0.005	2.86	0.020	3.35	0.034	3.80	0.044	4.14	0.051	4.39

Table 8: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ for $p = 1$ and $\alpha = 0.1$.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
2	0.176	10.85	0.637	15.10	1.086	19.25	1.513	23.18	1.950	27.22
3	0.155	10.82	0.523	14.68	0.838	18.06	1.110	21.01	1.371	23.89
4	0.138	10.79	0.447	14.41	0.690	17.36	0.889	19.81	1.072	22.10
5	0.125	10.78	0.392	14.21	0.590	16.87	0.746	19.01	0.887	20.97
6	0.114	10.76	0.349	14.06	0.518	16.53	0.646	18.45	0.759	20.17
7	0.105	10.75	0.316	13.94	0.462	16.26	0.570	18.03	0.664	19.59
8	0.097	10.74	0.288	13.85	0.418	16.05	0.512	17.69	0.592	19.12
9	0.091	10.73	0.266	13.76	0.381	15.87	0.464	17.42	0.535	18.75
10	0.085	10.72	0.246	13.70	0.351	15.72	0.426	17.20	0.488	18.44
11	0.080	10.71	0.230	13.64	0.326	15.60	0.393	17.01	0.449	18.19
12	0.076	10.71	0.215	13.58	0.304	15.49	0.365	16.84	0.416	17.97
13	0.072	10.70	0.203	13.53	0.285	15.39	0.341	16.70	0.387	17.78
14	0.068	10.70	0.191	13.49	0.268	15.31	0.321	16.58	0.363	17.62
15	0.065	10.69	0.181	13.45	0.253	15.24	0.302	16.47	0.341	17.48
16	0.062	10.69	0.172	13.42	0.240	15.17	0.286	16.37	0.322	17.35
17	0.059	10.69	0.164	13.39	0.228	15.11	0.271	16.29	0.305	17.24
18	0.057	10.68	0.157	13.37	0.218	15.06	0.258	16.22	0.290	17.15
19	0.054	10.68	0.150	13.34	0.208	15.01	0.246	16.15	0.276	17.05
20	0.052	10.67	0.144	13.32	0.199	14.97	0.235	16.09	0.264	16.97
21	0.050	10.67	0.138	13.30	0.191	14.93	0.225	16.03	0.252	16.90
22	0.049	10.67	0.133	13.28	0.183	14.90	0.216	15.98	0.242	16.83
23	0.047	10.67	0.128	13.27	0.176	14.87	0.208	15.93	0.232	16.77
24	0.045	10.67	0.124	13.25	0.170	14.84	0.200	15.88	0.224	16.71
25	0.044	10.66	0.120	13.23	0.164	14.80	0.193	15.84	0.215	16.66
26	0.043	10.66	0.116	13.22	0.159	14.78	0.186	15.81	0.208	16.60
27	0.041	10.66	0.112	13.20	0.153	14.75	0.180	15.77	0.201	16.56
28	0.040	10.66	0.109	13.19	0.149	14.73	0.174	15.74	0.194	16.52
29	0.039	10.65	0.105	13.18	0.144	14.71	0.169	15.71	0.188	16.48
30	0.038	10.65	0.102	13.17	0.140	14.69	0.164	15.68	0.182	16.45

Table 9: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ for $p = 2$ and $\alpha = 0.01$.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
2	0.065	6.38	0.301	7.80	0.579	9.46	0.861	11.16	1.160	12.95
3	0.055	6.38	0.243	7.72	0.441	9.16	0.623	10.51	0.801	11.85
4	0.047	6.38	0.204	7.66	0.359	8.98	0.494	10.15	0.620	11.27
5	0.042	6.37	0.177	7.63	0.305	8.86	0.412	9.92	0.509	10.90
6	0.038	6.37	0.156	7.60	0.265	8.77	0.354	9.76	0.433	10.65
7	0.034	6.37	0.140	7.58	0.235	8.71	0.311	9.64	0.378	10.47
8	0.031	6.37	0.126	7.56	0.212	8.66	0.278	9.54	0.336	10.32
9	0.029	6.37	0.116	7.54	0.192	8.61	0.252	9.47	0.302	10.21
10	0.027	6.37	0.106	7.53	0.176	8.58	0.230	9.41	0.275	10.12
11	0.025	6.37	0.099	7.52	0.163	8.55	0.211	9.35	0.252	10.04
12	0.023	6.37	0.092	7.51	0.151	8.52	0.196	9.31	0.233	9.98
13	0.022	6.37	0.086	7.50	0.141	8.50	0.183	9.27	0.217	9.92
14	0.021	6.37	0.081	7.49	0.133	8.48	0.171	9.24	0.203	9.87
15	0.019	6.37	0.077	7.48	0.125	8.47	0.161	9.21	0.190	9.83
16	0.018	6.37	0.073	7.48	0.118	8.45	0.152	9.19	0.180	9.80
17	0.018	6.37	0.069	7.47	0.112	8.44	0.144	9.16	0.170	9.76
18	0.017	6.37	0.066	7.47	0.107	8.42	0.137	9.14	0.161	9.74
19	0.016	6.36	0.063	7.46	0.102	8.41	0.130	9.13	0.153	9.71
20	0.015	6.36	0.060	7.46	0.097	8.40	0.124	9.11	0.146	9.69
21	0.015	6.36	0.057	7.46	0.093	8.39	0.119	9.09	0.140	9.66
22	0.014	6.36	0.055	7.45	0.089	8.38	0.114	9.08	0.134	9.64
23	0.014	6.36	0.053	7.45	0.086	8.38	0.109	9.07	0.128	9.63
24	0.013	6.36	0.051	7.45	0.082	8.37	0.105	9.06	0.123	9.61
25	0.013	6.36	0.049	7.45	0.079	8.36	0.101	9.05	0.119	9.59
26	0.012	6.36	0.048	7.44	0.077	8.36	0.098	9.03	0.115	9.58
27	0.012	6.36	0.046	7.44	0.074	8.35	0.094	9.02	0.111	9.57
28	0.011	6.36	0.044	7.44	0.072	8.35	0.091	9.02	0.107	9.55
29	0.011	6.36	0.043	7.44	0.069	8.34	0.088	9.01	0.103	9.54
30	0.011	6.36	0.042	7.44	0.067	8.34	0.086	9.00	0.100	9.53

Table 10: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ for $p = 2$ and $\alpha = 0.05$.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
2	0.043	4.80	0.214	5.59	0.431	6.59	0.660	7.65	0.911	8.80
3	0.035	4.80	0.166	5.55	0.318	6.43	0.463	7.29	0.608	8.17
4	0.030	4.80	0.137	5.52	0.254	6.34	0.361	7.11	0.463	7.85
5	0.026	4.80	0.116	5.51	0.213	6.29	0.297	6.99	0.376	7.65
6	0.023	4.80	0.101	5.50	0.183	6.25	0.253	6.90	0.317	7.52
7	0.020	4.80	0.090	5.49	0.161	6.21	0.221	6.84	0.275	7.42
8	0.018	4.80	0.081	5.48	0.144	6.19	0.197	6.80	0.243	7.35
9	0.017	4.80	0.073	5.47	0.130	6.17	0.177	6.76	0.218	7.29
10	0.015	4.80	0.067	5.47	0.119	6.15	0.161	6.73	0.197	7.24
11	0.014	4.80	0.062	5.47	0.109	6.14	0.148	6.71	0.181	7.20
12	0.013	4.80	0.058	5.46	0.101	6.13	0.136	6.69	0.167	7.17
13	0.012	4.80	0.054	5.46	0.094	6.12	0.127	6.67	0.155	7.14
14	0.012	4.80	0.050	5.46	0.088	6.11	0.118	6.65	0.144	7.12
15	0.011	4.80	0.047	5.45	0.083	6.11	0.111	6.64	0.135	7.10
16	0.010	4.80	0.045	5.45	0.078	6.10	0.105	6.63	0.127	7.08
17	0.010	4.80	0.042	5.45	0.074	6.09	0.099	6.62	0.120	7.07
18	0.009	4.80	0.040	5.45	0.070	6.09	0.094	6.61	0.114	7.05
19	0.009	4.80	0.038	5.45	0.067	6.08	0.089	6.60	0.108	7.04
20	0.008	4.80	0.037	5.45	0.064	6.08	0.085	6.59	0.103	7.03
21	0.008	4.80	0.035	5.44	0.061	6.08	0.081	6.59	0.098	7.02
22	0.008	4.80	0.034	5.44	0.058	6.07	0.078	6.58	0.094	7.01
23	0.007	4.80	0.032	5.44	0.056	6.07	0.075	6.57	0.090	7.00
24	0.007	4.80	0.031	5.44	0.054	6.07	0.072	6.57	0.087	6.99
25	0.007	4.80	0.030	5.44	0.052	6.06	0.069	6.56	0.083	6.99
26	0.007	4.80	0.029	5.44	0.050	6.06	0.066	6.56	0.080	6.98
27	0.006	4.80	0.028	5.44	0.048	6.06	0.064	6.55	0.077	6.97
28	0.006	4.80	0.027	5.44	0.046	6.06	0.062	6.55	0.075	6.97
29	0.006	4.80	0.026	5.44	0.045	6.05	0.060	6.55	0.072	6.96
30	0.006	4.80	0.025	5.44	0.044	6.05	0.058	6.54	0.070	6.96

Table 11: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ for $p = 2$ and $\alpha = 0.1$.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
3	0.153	13.08	0.531	17.38	0.880	21.34	1.196	24.92	1.506	28.44
4	0.138	13.06	0.459	17.10	0.732	20.58	0.965	23.58	1.185	26.42
5	0.126	13.04	0.406	16.89	0.630	20.05	0.814	22.68	0.983	25.11
6	0.116	13.03	0.364	16.73	0.555	19.66	0.707	22.03	0.843	24.17
7	0.107	13.01	0.331	16.60	0.497	19.35	0.626	21.53	0.740	23.48
8	0.100	13.00	0.303	16.49	0.451	19.11	0.562	21.13	0.660	22.92
9	0.093	12.99	0.280	16.41	0.412	18.91	0.512	20.83	0.596	22.49
10	0.088	12.98	0.260	16.33	0.381	18.73	0.469	20.55	0.545	22.12
11	0.083	12.98	0.243	16.26	0.354	18.59	0.434	20.33	0.501	21.81
12	0.079	12.97	0.229	16.20	0.330	18.47	0.404	20.14	0.465	21.55
13	0.075	12.97	0.216	16.15	0.310	18.36	0.377	19.97	0.433	21.33
14	0.071	12.96	0.204	16.11	0.292	18.26	0.354	19.83	0.406	21.14
15	0.068	12.96	0.194	16.06	0.276	18.18	0.334	19.70	0.382	20.96
16	0.065	12.95	0.184	16.03	0.262	18.10	0.316	19.58	0.361	20.80
17	0.062	12.95	0.176	15.99	0.249	18.03	0.300	19.48	0.342	20.66
18	0.059	12.94	0.168	15.96	0.238	17.96	0.286	19.38	0.325	20.54
19	0.057	12.94	0.161	15.93	0.227	17.91	0.273	19.29	0.309	20.43
20	0.055	12.94	0.155	15.91	0.217	17.86	0.261	19.22	0.295	20.33
21	0.053	12.93	0.149	15.89	0.209	17.81	0.250	19.15	0.283	20.23
22	0.051	12.93	0.143	15.86	0.201	17.77	0.240	19.09	0.271	20.15
23	0.050	12.93	0.138	15.84	0.193	17.72	0.231	19.03	0.260	20.07
24	0.048	12.92	0.133	15.82	0.186	17.68	0.222	18.97	0.251	20.00
25	0.047	12.92	0.129	15.80	0.180	17.65	0.214	18.92	0.241	19.93
26	0.045	12.92	0.125	15.79	0.174	17.62	0.207	18.87	0.233	19.87
27	0.044	12.92	0.121	15.77	0.168	17.58	0.200	18.83	0.225	19.81
28	0.043	12.92	0.117	15.76	0.163	17.55	0.194	18.79	0.218	19.76
29	0.041	12.92	0.114	15.74	0.158	17.53	0.188	18.75	0.211	19.71
30	0.040	12.91	0.111	15.73	0.153	17.50	0.182	18.71	0.204	19.66

Table 12: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha;\tilde{a}(\gamma),k,p)$ for $p=3$ and $\alpha=0.01$.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
3	0.055	8.24	0.250	9.77	0.470	11.49	0.683	13.16	0.902	14.86
4	0.048	8.24	0.213	9.71	0.387	11.29	0.547	12.74	0.702	14.17
5	0.043	8.24	0.186	9.67	0.331	11.15	0.459	12.47	0.579	13.73
6	0.039	8.24	0.165	9.64	0.289	11.04	0.396	12.27	0.494	13.42
7	0.035	8.24	0.149	9.61	0.258	10.96	0.349	12.13	0.432	13.19
8	0.033	8.23	0.135	9.59	0.232	10.90	0.312	12.01	0.384	13.01
9	0.030	8.23	0.124	9.57	0.212	10.85	0.283	11.91	0.346	12.87
10	0.028	8.23	0.115	9.56	0.195	10.81	0.259	11.84	0.315	12.74
11	0.026	8.23	0.107	9.55	0.180	10.77	0.239	11.77	0.289	12.65
12	0.025	8.23	0.100	9.54	0.168	10.74	0.221	11.71	0.267	12.56
13	0.023	8.23	0.094	9.52	0.157	10.71	0.207	11.67	0.249	12.49
14	0.022	8.23	0.088	9.52	0.147	10.69	0.194	11.62	0.233	12.43
15	0.021	8.23	0.084	9.51	0.139	10.67	0.182	11.58	0.219	12.37
16	0.020	8.23	0.079	9.50	0.132	10.65	0.172	11.55	0.206	12.32
17	0.019	8.23	0.076	9.50	0.125	10.63	0.163	11.52	0.195	12.28
18	0.018	8.23	0.072	9.49	0.119	10.62	0.155	11.50	0.185	12.25
19	0.017	8.23	0.069	9.49	0.113	10.60	0.148	11.47	0.176	12.21
20	0.017	8.23	0.066	9.48	0.108	10.59	0.141	11.45	0.168	12.17
21	0.016	8.23	0.063	9.48	0.104	10.58	0.135	11.43	0.160	12.15
22	0.015	8.23	0.061	9.47	0.100	10.56	0.129	11.41	0.154	12.12
23	0.015	8.23	0.058	9.47	0.096	10.56	0.124	11.40	0.148	12.10
24	0.014	8.23	0.056	9.47	0.092	10.55	0.119	11.38	0.142	12.07
25	0.014	8.23	0.054	9.46	0.089	10.54	0.115	11.37	0.137	12.05
26	0.013	8.23	0.053	9.46	0.086	10.53	0.111	11.35	0.132	12.04
27	0.013	8.23	0.051	9.46	0.083	10.53	0.107	11.34	0.127	12.02
28	0.013	8.23	0.049	9.46	0.080	10.52	0.104	11.33	0.123	12.00
29	0.012	8.23	0.048	9.45	0.078	10.51	0.100	11.32	0.119	11.99
30	0.012	8.23	0.046	9.45	0.075	10.51	0.097	11.31	0.115	11.97

Table 13: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ for $p = 3$ and $\alpha = 0.05$.

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$		$\gamma = 0.15$		$\gamma = 0.2$	
k	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val	$\tilde{a}(\gamma)$	Crit Val
3	0.036	6.48	0.176	7.35	0.347	8.42	0.522	9.51	0.705	10.66
4	0.031	6.48	0.147	7.32	0.281	8.32	0.410	9.28	0.539	10.26
5	0.027	6.48	0.126	7.31	0.237	8.25	0.339	9.14	0.439	10.01
6	0.024	6.48	0.111	7.29	0.205	8.20	0.290	9.03	0.371	9.84
7	0.022	6.47	0.099	7.28	0.181	8.16	0.254	8.96	0.322	9.71
8	0.020	6.47	0.089	7.27	0.162	8.13	0.226	8.90	0.285	9.61
9	0.018	6.47	0.081	7.26	0.147	8.11	0.204	8.85	0.255	9.53
10	0.017	6.47	0.074	7.26	0.134	8.08	0.186	8.81	0.232	9.47
11	0.015	6.47	0.069	7.25	0.124	8.07	0.170	8.77	0.212	9.41
12	0.014	6.47	0.064	7.25	0.115	8.05	0.158	8.75	0.196	9.37
13	0.013	6.47	0.060	7.24	0.107	8.04	0.147	8.72	0.182	9.33
14	0.013	6.47	0.056	7.24	0.100	8.03	0.137	8.70	0.169	9.30
15	0.012	6.47	0.053	7.24	0.094	8.02	0.129	8.68	0.159	9.27
16	0.011	6.47	0.050	7.23	0.089	8.01	0.121	8.66	0.149	9.25
17	0.011	6.47	0.048	7.23	0.084	8.00	0.115	8.65	0.141	9.22
18	0.010	6.47	0.045	7.23	0.080	7.99	0.109	8.64	0.134	9.20
19	0.010	6.47	0.043	7.23	0.076	7.99	0.103	8.62	0.127	9.18
20	0.009	6.47	0.041	7.23	0.073	7.98	0.099	8.61	0.121	9.17
21	0.009	6.47	0.040	7.22	0.070	7.98	0.094	8.60	0.116	9.15
22	0.009	6.47	0.038	7.22	0.067	7.97	0.090	8.59	0.111	9.14
23	0.008	6.47	0.036	7.22	0.064	7.97	0.087	8.59	0.106	9.13
24	0.008	6.47	0.035	7.22	0.062	7.96	0.083	8.58	0.102	9.11
25	0.008	6.47	0.034	7.22	0.059	7.96	0.080	8.57	0.098	9.10
26	0.007	6.47	0.033	7.22	0.057	7.96	0.077	8.57	0.094	9.09
27	0.007	6.47	0.032	7.22	0.055	7.95	0.075	8.56	0.091	9.08
28	0.007	6.47	0.030	7.22	0.053	7.95	0.072	8.55	0.088	9.08
29	0.007	6.47	0.030	7.21	0.052	7.95	0.070	8.55	0.085	9.07
30	0.007	6.47	0.029	7.21	0.050	7.94	0.067	8.54	0.082	9.06

Table 14: Weights $\tilde{a}(\gamma)$ and critical values $\tilde{F}^{-1}(1-\alpha; \tilde{a}(\gamma), k, p)$ for $p = 3$ and $\alpha = 0.1$.

Like the confidence sets $CS_{R,P}$ and CS_R discussed in the main text, JK confidence sets are based on the fact that, evaluated at the true parameter value, $K_{\Omega,f}(\theta_0)$ and $S(\theta_0) - K_{\Omega,f}(\theta_0)$ are asymptotically independent and distributed χ^2_p and χ^2_{k-p}. To construct the JK confidence set for β, define $J_{\Omega,f}(\theta) = S(\theta) - K_{\Omega,f}(\theta)$ and for constants α_K, $\alpha_J \in (0,1)$ let

$$CS_{JK} = \left\{\beta : \exists \theta \, s.t. \, \beta = f(\theta), \, K_{\Omega,f}(\theta) \leq \chi^2_{1-\alpha_K,p}, \text{ and } J_{\Omega,f}(\theta) \leq \chi^2_{1-\alpha_J,k-p}\right\}$$

or, equivalently

$$CS_{JK} = \left\{\beta : \min_{\theta:\beta=f(\theta)} \max\left\{K_{\Omega,f}(\theta) - \chi^2_{1-\alpha_K,p}, J_{\Omega,f}(\theta) - \chi^2_{1-\alpha_J,k-p}\right\} \leq 0\right\}.$$

Under Assumptions 2-6 one can show using the proofs of Theorems 2 and 3 that the asymptotic coverage of this robust confidence set is at least $(1-\alpha_J)(1-\alpha_K) = 1-\alpha_J-\alpha_K+\alpha_J\alpha_K$ under both weak and strong identification.

To construct a preliminary confidence set $CS_{JK,P}$ such that $CS_{R,P} = CS_{JK,P}$ will satisfy Assumption 1, let $CS_{JK,P}$ be the JK confidence set based on $(\alpha_{J,P}, \alpha_{K,P})$ satisfying $(1-\alpha_{J,P})(1-\alpha_{K,P}) = 1-\alpha-\gamma$ and $\alpha_{K,P} > \alpha$. $CS_{JK,P}$ will have sequential coverage at least $1-\alpha-\gamma$ under weak identification,

$$SCP(CS_{JK,P}, \Xi_W) \geq 1-\alpha-\gamma$$

and will be contained in the non-robust confidence set with probability tending to one under strong identification

$$\inf_{\xi \in \Xi_S} Pr_{T,\xi}\{CS_{JK,P} \subseteq CS_{NR}\} \to 1$$

for CS_{NR} as in (13). Thus, this choice satisfies Assumption 1(1) and (3). If we define $CS_R = CS_{JK}$ for $\alpha_K \leq \alpha_{K,P}$, $\alpha_J \leq \alpha_{J,P}$, and $(1-\alpha_J)(1-\alpha_K) = 1-\alpha$, we can see that $CS_{JK,P} \subseteq CS_{JK}$ for all realizations of the data, so Assumption 1(2) is satisfied as well. Thus, under the assumptions of Theorem 3 the choice $CS_{R,P} = CS_{JK,P}$ and $CS_R = CS_{JK}$ satisfies Assumption 1 and thus yields two-step confidence sets which control coverage distortions. Simulation results based on the same calibrations studied in Section 4 (available on request)

show that two-step procedures based on JK confidence sets with $\alpha_K = 0.8 \cdot \alpha$ and $\alpha_{K,P} = 0.8\,(\alpha + \gamma)$ perform comparably to the approach developed in the main text for $\alpha = 5\%$ and $\gamma = 10\%$.