Robust Inference on Macro Equations with Shock Proxies[*]

Ryohei Oishi[†]

March 2026

Abstract

This paper proposes weak-instrument-robust inference methods that achieve accurate size in finite samples for macroeconomic equations identified by structural shock proxies. By utilizing a restricted heteroskedasticity- and autocorrelation-robust (HAR) long-run variance, the methods control the size of the Anderson-Rubin and Kleibergen's LM test statistics with Almon-parameterized instruments. Applying these methods to the Phillips curve highlights the empirical importance of size accuracy: unlike existing methods that may introduce size distortion, the proposed methods do not reject either a completely flat or a very steep Phillips curve. Turning to fiscal policy rules, I find that the tax multiplier is statistically significantly positive and approximately one.

JEL Classification: C12, C26, C32, E31, E62

Keywords: Weak instruments, HAR inference, Shock proxy, Phillips curve, Fiscal multiplier

[*] I am sincerely grateful to my supervisors, Vincent Sterk and Morten O. Ravn, for their helpful comments and support, and to Daniel J. Lewis for his guidance throughout the project. I thank seminar participants at UCL and York for their helpful comments and discussions.

[†] University College London, Gower Street, London, WC1E 6BT, UK (e-mail: ryohei.oishi.22@ucl.ac.uk)

1 Introduction

Econometric analysis of structural macroeconomic models has been challenging due to endogeneity stemming from measurement error and simultaneity. One of the most widely adopted approaches in the applied literature is to use predetermined macroeconomic variables as instruments (Hansen and Singleton, 1982; Galí and Gertler, 1999); however, this relies on stringent assumptions, such as the absence of serial correlation in the disturbances and measurement errors. Furthermore, the correlation between the predetermined and endogenous variables is often low, causing the (many) weak-instrument problem.[1]

To address these challenges, Barnichon and Mesters (2020) propose using a sequence of structural shock proxies with Almon parameterization (Almon, 1965) as instruments, and conducting inference using the Anderson and Rubin (1949, henceforth AR) statistic, which is robust to weak instruments. The externally identified shock proxies, such as high-frequency identified monetary policy surprises (e.g., Gürkaynak, Sack, and Swanson, 2005), are arguably more likely to be exogenous than lagged macroeconomic variables. Moreover, Almon parameterization mitigates the many-instrument problem by compressing the information from a sequence of shocks, typically 12 to 20 lags for quarterly data, into a few instruments.[2]

Despite the substantial progress made by Barnichon and Mesters (2020), the econometric toolkit remains far from complete. Particularly, we lack a robust inference procedure suitable for *overidentified* models with Almon-parameterized instruments. This is because, first, overidentification complicates the interpretation of AR confidence sets. Because the AR statistic jointly tests the parameters and the overidentifying restrictions, a narrow confidence set may simply reflect a violation of overidentifying restrictions rather than precise parameter estimates. Second, the degrees of freedom of its limiting distribution depend on the number of instruments rather than the number of structural parameters. Furthermore, as demonstrated in this paper, the method of Barnichon and

[1] See, e.g., Andrews, Stock, and Sun (2019); Mikusheva (2021) for recent survey.

[2] This choice of lags is related to the interpretation of this procedure as a "regression in impulse response space," which is discussed in Section 2, and based on the observation that the impulse responses of typical macroeconomic variables are different from zero for three to five years.

Mesters (2020) tends to over-reject the null hypothesis in finite samples.

This paper proposes a new weak-instrument-robust test that is particularly useful for overidentified models: the *Almon-parameterized* Kleibergen (2002) Lagrange multiplier (henceforth KLM) statistic. It overcomes the disadvantages of the Almon-parameterized AR test for overidentified models because its limiting distribution has degrees of freedom equal to the number of structural parameters, and the statistic remains pivotal even when the instruments are weak or irrelevant. Therefore, when the model is overidentified, one can construct a robust confidence set by inverting the Almon-parameterized KLM statistic.

Constructing the local-to-zero weak-instrument asymptotic framework (Staiger and Stock, 1997) for the Almon-parameterized KLM statistic is non-trivial. This is because, as in Barnichon and Mesters (2020), the Almon parameterization induces non-stationary instruments whose variances diverge at different rates by taking weighted partial sums of the shock sequence with polynomial weights. In this paper, I construct the weak-instrument asymptotic framework by applying a different scaling factor to the first-stage parameters of each Almon-parameterized instrument, rather than the standard uniform scaling by $\sqrt{n}$, where n is the sample size. Intuitively, because the Almon parameterization artificially inflates the signal, we must adapt these tailored scalings to recover the original information content. I show that this framework is indeed informationally equivalent to the standard weak-instrument asymptotics of Staiger and Stock (1997).[3]

The key idea for controlling size accurately in finite samples for both the Almon-parameterized AR and KLM statistics is to apply the *restricted* heteroskedasticity- and autocorrelation-robust (HAR) long-run variance estimator computed by imposing the null hypothesis. While its use was initially proposed by Lazarus, Lewis, Stock, and Watson (2018) for hypothesis testing in a time-series regression model, the same remedy works for weak-instrument-robust inference, as confirmed by the Monte Carlo simulations in this paper. In contrast, when the *unrestricted* long-run variance is used, which is

[3] In short, these different scalings are necessary to ensure that the concentration parameter converges, rather than diverges, in the limit under weak-instrument asymptotics.

the approach taken by Barnichon and Mesters (2020), one may encounter over-rejection in finite samples. This finding is consistent with the well-established fact in the HAR literature that conventional hypothesis testing tends to over-reject the null hypothesis in finite samples (Müller, 2014; Lazarus et al., 2018).

Implementing the Almon-parameterized AR and KLM statistics with the restricted long-run variance estimator is computationally straightforward. Because the restricted residual simply equals the structural equation error in this case, one can directly apply standard HAR long-run variance calculations. Consequently, only a minor modification to the replication code of Barnichon and Mesters (2020) is needed to employ the Almon-parameterized AR statistic with the restricted long-run variance. To facilitate applied work, example code that includes both the Almon-parameterized AR and KLM statistics is available online.[4]

A natural question from an applied perspective is whether such finite-sample size control matters in practice. The answer is yes: I show that the empirical conclusion regarding the shape of the New Keynesian Phillips curve (henceforth, the Phillips curve) varies depending on whether the size-distorted or size-accurate methods are used. When the size-distorted methods are applied, one may conclude that the slope of the Phillips curve in the US is significantly different from zero (Barnichon and Mesters, 2020). In contrast, the size-accurate methods imply that neither a completely flat nor a very steep Phillips curve is rejected, suggesting that the seemingly significant previous results are mainly driven by over-rejection.

Given the effectiveness of the proposed methods, I apply them to fiscal policy rules and find that the tax multiplier is statistically significantly positive and approximately one. Unlike the Phillips curve application, where the size-accurate robust inference method overturns the previous findings, this application demonstrates that robust methods still deliver statistically significant results. More concretely, I measure the tax and spending multipliers by building on Caldara and Kamps (2017), who establish a one-to-one mapping between the systematic responses of fiscal policy to output and the fiscal mul-

[4] `https://github.com/ryohei-oishi/Robust-Inference-Shock-Proxy-Example-Public`.

tipliers. Hence, I first construct robust confidence sets for the tax and spending rule parameters, and then translate these into multiplier estimates. To my knowledge, this exercise provides the first evidence of a statistically significant tax multiplier, accounting for weak-instrument issues. The spending multiplier, by contrast, is not statistically significant, although the result suggests it is likely moderately above one.

Related literature. First and foremost, this paper contributes to the vast literature on weak-instrument-robust inference for structural macroeconomic equations via a limited-information approach.[5] The most closely related paper is Barnichon and Mesters (2020), which proposed weak-instrument robust inference using a sequence of structural shock proxies as instruments and showed the equivalence of this procedure as a "regression in impulse response space." The key distinction from Barnichon and Mesters (2020) lies in finite-sample size accuracy. The methods proposed in this paper maintain correct size, whereas theirs exhibit finite-sample size distortions. I also propose the Almon-parameterized KLM statistic, which overcomes some of the AR statistic's disadvantages. Moreover, I demonstrate that their empirical finding of a statistically significant Phillips curve slope is largely driven by these size distortions.[6] Furthermore, the idea of using a "regression in impulse response space" to estimate the object of macroeconomic interest, in my case the structural parameters, is also applied in McKay and Wolf (2023); Barnichon and Mesters (2023) for policy counterfactual analysis.

Second, this paper relates to the HAR inference literature, particularly studies that address finite-sample size distortions in the standard nonparametric approach of Newey and West (1987) and Andrews (1991). Lazarus et al. (2018) survey the recent literature and argue that the conventional nonparametric approach under small-b asymptotics leads to over-rejection in regression models. I show that the same type of size distortion occurs

[5] See, for example, Yogo (2004); Ascari, Magnusson, and Mavroeidis (2021); Haque and Magnusson (2023); Ascari, Haque, Magnusson, and Mavroeidis (2024) for the Euler equation, Kleibergen and Mavroeidis (2009); Magnusson and Mavroeidis (2014); Dovì (2025); Inoue, Rossi, and Wang (2025) for the Phillips curve, and Mavroeidis (2010); Carlevaro, Haque, and Magnusson (2025) for the Taylor rule. Also, see Lewis and Mertens (2024). Although the empirical applications of this paper focus on the Phillips curve and fiscal rules, the method can be applied to other structural macroeconomic models.

[6] See Mavroeidis, Plagborg-Møller, and Stock (2014) and Furlanetto and Lepetit (2025) for surveys of the Phillips curve estimation.

in weak-instrument-robust inference. Furthermore, the proposed use of the restricted long-run variance builds on Lazarus et al. (2018) and Vogelsang (2018), and the robustness analysis using the fixed-b critical values is motivated by works such as Kiefer and Vogelsang (2002a,b, 2005); Jansson (2004); Sun, Phillips, and Jin (2008); Sun (2014).

Third, the empirical analysis contributes to the literature on fiscal policy rules and multipliers.[7] Methodologically, this paper complements the popular proxy structural VAR (SVAR) approach (e.g., Mertens and Ravn, 2013, 2014; Caldara and Kamps, 2017; Angelini, Caggiano, Castelnuovo, and Fanelli, 2023; Gregory, McNeil, and Smith, 2024; Keweloh, Klein, and Prüser, 2025). Among these studies, this paper is most closely related to Caldara and Kamps (2017). They identify a proxy SVAR using proxies *uncorrelated* with the fiscal shock and measure the fiscal multipliers. Although I use the same instruments, my approach differs methodologically from theirs: I apply the weak-instrument-robust method to a fiscal policy rule, whereas they assume strong identification of a structural VAR. Empirically, I provide the first weak-instrument-robust evidence that the tax multiplier is positive and statistically significant.[8]

Structure of the paper. Section 2 introduces the model setup, presents the Almon-parameterized AR and KLM statistics and their limiting distributions, and proposes a method to control finite-sample size accurately. Section 3 validates the finite-sample size accuracy of the proposed methods via Monte Carlo simulations. Section 4 highlights the practical importance of the size-accurate method through an empirical analysis of the Phillips curve. Section 5 examines fiscal policy rules and multipliers. Section 6 concludes. The Supplementary Appendix includes technical assumptions, the proof of the theorem, additional simulations and empirical results, and robustness checks.

[7] See, Ramey (2016, 2019); Varela and Ribeiro (2025) for surveys.

[8] In a slightly different context, Mertens and Montiel Olea (2018) study short-run tax elasticities of income using exogenous variations in marginal tax rates, taking into account weak instruments, but do not directly estimate fiscal multipliers. Hebous and Zimmermann (2018) is an exceptional case that applies the conditional likelihood test of Moreira (2003) to instrumental-variable local projections, and they find that the impulse response of output growth is not significantly different from zero.

Notation. Throughout the paper, $\xrightarrow{p}$ and $\Rightarrow$ denote convergence in probability and in distribution, respectively; $\text{vec}(A)$ denotes a column vectorization of a matrix A; $P_A = A(A'A)^{-1}A'$ is a projection matrix onto the column of the full-rank matrix A; $M_A = I - P_A$ is a projection onto a space orthogonal to A; and $\otimes$ denotes the Kronecker product.

2 Weak-instrument-robust inference with HAR size control

2.1 Linear IV model with Almon-parameterized shock proxies

The structural equation. The setup is a standard linear model with endogenous regressors:

$$y_t = w_t'\delta + u_t, \quad E[w_t u_t] \neq 0, \tag{1}$$

where we are interested in the inference on the structural parameter δ.

The model (1) nests various linear(ized) structural macroeconomic equations. For example, consider the Phillips curve: $\pi_t = \gamma_f E_t[\pi_{t+1}] + \lambda x_t + \varepsilon_t^c$, where π_t is inflation, $E_t[\pi_{t+1}]$ is inflation expectations, x_t is the output gap, and ε_t^c is a cost-push disturbance. Since inflation expectations and the output gap are not directly observable, they must be replaced with some observables, e.g., the realized inflation π_{t+1} and a measure of the output gap x_t^o. Then, the Phillips curve can be written as equation (1) where $y_t = \pi_t$, $w_t = [\pi_{t+1}, x_t^o]'$, $\delta = [\gamma_f, \lambda]'$ and $u_t = \varepsilon_t^c - \gamma_f(\pi_{t+1} - E_t[\pi_{t+1}]) - \lambda(x_t^o - x_t)$.

The endogeneity $E[w_t u_t] \neq 0$ is typically due to measurement errors and simultaneity in a macroeconomic setup. As is evident in w_t and u_t above for the Phillips curve, replacing inflation expectations and the output gap with observables induces endogeneity whenever there is measurement error. Furthermore, the cost-push disturbance is correlated with the output gap due to endogenous responses by a monetary authority and households: when a positive cost-push disturbance exogenously increases inflation, the central bank raises interest rates, households reduce consumption, and output is then reduced.

One of the most popular approaches to estimating (1) in the applied macroeconomic literature is to use predetermined variables as instruments, under the assumption of rational expectations and the absence of serial correlation in measurement errors and disturbances. However, the latter conditions are rarely satisfied in practice. For example, in a standard quantitative DSGE model (e.g., Smets and Wouters, 2007), the disturbances are modeled as autoregressive and/or moving-average processes. Moreover, the correlation between the predetermined and endogenous variables is often low, leading to the weak-instrument problem.[9]

Instruments: the Almon-parameterized shock proxies. As argued in Barnichon and Mesters (2020), a sequence of externally identified structural shock proxies is more likely to be valid instruments than the predetermined variables, since it allows measurement errors and disturbances to be autocorrelated. Formally, suppose there exists a sequence of structural shock proxies $\xi_{t:t-H} = (\xi_t, \cdots, \xi_{t-H})'$ where H is the number of lags. While ξ_t can be a vector when multiple types of shock proxies are available, for ease of notation, I assume it is a scalar.[10] I assume that the proxy satisfies the exogeneity condition $E[u_t \xi_s] = 0$ for all t and s. Since the relevance condition is related to weak-instrument robust tests, I discuss it in the next section.

In practice, one needs to choose the lags H, and typical macroeconomic applications suggest that using $H = 12$ to $H = 20$ lags for quarterly data is informative for identification. This choice of H is based on the interpretation of identification using the sequence of shock proxies as a "regression in impulse response space". Specifically, Barnichon and Mesters (2020) show that identifying δ by the shock sequence is equivalent to regressing the impulse responses of y to the shock on those of w in the model (1). Since typical

[9] Figures 4 and 5 of Mavroeidis et al. (2014) effectively demonstrate the empirical difficulties of this popular approach. They estimate more than 600,000 Phillips curve specifications by varying datasets, sample periods, instruments, and so forth, and show that the resulting point estimates exhibit substantial dispersion, even with signs opposite to theoretical predictions.

[10] For instance, one could utilize multiple types of monetary surprises, such as the measures of forward guidance and large-scale asset purchase shocks constructed by Swanson (2021). Investigating the optimal combination of multiple proxies remains an interesting avenue for future research.

macroeconomic impulse responses are nonzero for between 3 and 5 years, the lags of $H = 12$ to $H = 20$ contain useful information for identification.

Although the "regression in impulse response space" interpretation suggests using $H \in [12, 20]$ lags of proxies as instruments for quarterly data, directly using them may lead to the many-instrument problem, especially when the number of endogenous variables is small. On the other hand, using a small H discards informative variation. To balance this tradeoff, following Barnichon and Mesters (2020), I apply the Almon parameterization to the sequence to construct a new set of instruments. Specifically, the Almon-parameterized instruments are defined as follows:

$$z_t = \left(\sum_{h=0}^{H} \xi_{t-h}, \sum_{h=0}^{H} h\xi_{t-h}, \sum_{h=0}^{H} h^2 \xi_{t-h} \right)', \tag{2}$$

and I assume that the endogenous variables are generated by

$$w_t = \Pi' z_t + v_t, \tag{3}$$

where Π is a full column rank matrix. The Almon parameterization can be interpreted as a shape restriction on the impulse responses. In particular, the Almon parameterization is equivalent to assuming that the impulse response is a quadratic polynomial, reflecting the common belief that macro impulse responses are smooth.[11]

Although the Almon parametrization (2) restricts the dimension of δ to be less than or equal to 3, this restriction is less stringent than it seems. This is because the number of

[11] Consider the distributed-lag regression of a variable x_t on the shock sequence $\xi_{t:t-H}$:

$$x_t = \theta^{(0)}\xi_t + \theta^{(1)}\xi_{t-1} + \cdots + \theta^{(h)}\xi_{t-h} + \cdots + \theta^{(H)}\xi_{t-H},$$

where the coefficients $(\theta^{(0)}, \cdots \theta^{(H)})'$ represent the impulse response function of x on ξ. Suppose that one imposes a shape restriction that the impulse response is quadratic: $\theta^{(h)} = a + bh + ch^2$. Then the regression can be represented as

$$\begin{aligned} x_t &= a\xi_t + (a + b + c)\xi_{t-1} + \cdots + (a + hb + h^2 c)\xi_{t-h} + \cdots + (a + Hb + H^2 c)\xi_{t-H} \\ &= a \sum_{h=0}^{H} \xi_{t-h} + b \sum_{h=0}^{H} h\xi_{t-h} + c \sum_{h=0}^{H} h^2 \xi_{t-h} = \theta_a' z_t, \end{aligned}$$

where $\theta_a = (a, b, c)$. Hence, the Almon-parametrization in (2) is equivalent to assuming that the impulse response is quadratic.

endogenous variables in a typical structural macroeconomic equation is small. Moreover, one can include higher-order Almon polynomials if the number of endogenous variables exceeds three or if these terms capture additional information by better approximating complex impulse responses.

Equations (1), (2), and (3) represent the data-generating process for the observations $\{y_t, w_t, z_t\}$ considered in this paper. For (ξ_t, u_t, v_t'), I put some technical assumptions following Barnichon and Mesters (2020), which are summarized as Assumption 1 in Supplementary Appendix A.

2.2 Weak-instrument-robust tests and HAR size control

Although the Almon parameterization mitigates the many-instrument problem, the weak-instrument issue may persist. It is well documented in the literature that structural shocks explain only a limited amount of variation in macroeconomic variables (e.g., Gorodnichenko and Lee, 2020; Plagborg-Møller and Wolf, 2022; Lewis and Mertens, 2024). Indeed, I show that the shock proxies used to identify the Phillips curve and fiscal policy rules in this paper are not sufficiently strong to support standard inference methods.

To conduct weak-instrument-robust inference on δ, I propose using the AR and KLM statistics with Almon-parameterized instruments for just-identified and overidentified models, respectively. Furthermore, to control size accurately in finite samples, I propose using the *restricted* HAR long-run variance computed by imposing the null hypothesis for both the AR and KLM statistics.

Almon-parameterized AR statistic. The Almon-parameterized AR test statistic with the *restricted* long-run variance estimator, denoted by AR_a^r, is defined as follows:

$$AR_a^r(\delta_0) = \hat{\theta}_a' \widetilde{\Sigma}_{\theta_a}^{-1} \hat{\theta}_a, \tag{4}$$

where

$$\hat{\theta}_a = \left(\sum_{t=H+1}^{n} z_t z_t'\right)^{-1} \left(\sum_{t=H+1}^{n} z_t(y_t - w_t'\delta_0)\right), \quad \widetilde{\Sigma}_{\theta_a} = \left(\sum_{t=H+1}^{n} z_t z_t'\right)^{-1} \tilde{s}_u^2,$$

$$\tilde{s}_u^2 = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} \widetilde{u}_t \widetilde{u}_s \kappa((t-s)/b_n), \tag{5}$$

$$\widetilde{u}_t = \widetilde{\widetilde{u}}_t - \bar{\widetilde{u}}, \quad \widetilde{\widetilde{u}}_t = y_t - w_t'\delta_0, \quad \bar{\widetilde{u}} = \frac{1}{n-H} \sum_{t=H+1}^{n} \widetilde{\widetilde{u}}_t, \tag{6}$$

$\kappa(\cdot)$ is a kernel satisfying standard assumptions summarized in Assumption 1 (v), and $\widetilde{u}_t$ is demeaned following Lazarus et al. (2018) and Vogelsang (2018).[12] Intuitively, the idea of the AR test relies on the premise that the regression coefficients of the error u_t on the instruments z_t should be zero in the following regression if the exogeneity condition holds:

$$u_t(\delta_0) := y_t - w_t'\delta_0 = \theta_a' z_t + \eta_t, \tag{7}$$

thus we can test $H_0 : \delta = \delta_0$ by equivalently testing $\theta_a = 0$.

The key idea for controlling size accurately in finite samples is to impose the null hypothesis when computing the long-run variance. Therefore, it is important to note that (5) with (6) imposes the null hypothesis $\theta_a = 0$; otherwise, the residual in (6) would be computed by subtracting $z_t'\hat{\theta}_a$, which leads to the unrestricted long-run variance estimator.

Accordingly, it is also possible to consider the Almon-parameterized AR test with the *unrestricted* long-run variance estimator, which is indeed the approach taken by Barnichon and Mesters (2020).[13] Specifically, their test statistic AR_a is defined similarly to AR_a^r in (4) but replaces the long-run variance estimator (5) and (6) with the unrestricted

[12] For both AR_a^r and AR_a (discussed next), the forms of $\widetilde{\Sigma}_{\theta_a}$ and $\widehat{\Sigma}_{\theta_a}$ are motivated by the asymptotic theory detailed in Supplementary Appendix B. Furthermore, for both cases, I use the Quadratic Spectral (QS) kernel for $\kappa(\cdot)$ and the bandwidth parameter set to $b_n = \lfloor 4((n-H)/100)^{2/9} \rfloor + 1$, following Barnichon and Mesters (2020), where $\lfloor x \rfloor$ is the largest integer less than or equal to x.

[13] Barnichon and Mesters (2020) mention in their Footnote 25 that the restricted and unrestricted long-run variance estimators are asymptotically equivalent, but they do not discuss their finite-sample properties.

version as follows:

$$\hat{s}_u^2 = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} \hat{u}_t \hat{u}_s \kappa((t-s)/b_n), \quad \hat{u}_t = (y_t - w_t'\delta_0) - z_t'\hat{\theta}_a. \tag{8}$$

Almon-parameterized KLM statistic. While the AR statistic has some desirable properties for the just-identified case, such as admissibility established by Chernozhukov, Hansen, and Jansson (2009), it has disadvantages when applied to overidentified models. For instance, the degrees of freedom of its limiting distribution exceed the number of structural parameters. Furthermore, because it tests the overidentifying restrictions simultaneously, the interpretation of the confidence sets is complicated: they may be narrow not because the parameter is estimated precisely, but because the overidentifying restrictions are rejected. To overcome these disadvantages, Kleibergen (2002) proposed the KLM statistic whose limiting distribution is $\chi^2(m)$ where m is the number of structural parameters, even when the instruments are weak. I next show that this property also holds for the *Almon-parameterized* KLM statistic.[14]

The Almon-parameterized KLM statistic with the restricted long-run variance, denoted $K_a^r(\delta_0)$, is defined as follows:

$$K_a^r(\delta_0) = \frac{(y - W\delta_0)' P_{\widetilde{W}(\delta_0)} (y - W\delta_0)}{\tilde{s}_u^2}, \tag{9}$$

where

$$\widetilde{W}(\delta_0) = Z\widetilde{\Pi}(\delta_0), \quad \widetilde{\Pi}(\delta_0) = (Z'Z)^{-1}Z'\left[W - (y - W\delta_0)\frac{\tilde{s}_{uw}}{\tilde{s}_u^2}\right],$$

$$\tilde{s}_{uw} = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} \widetilde{u}_t w_s \kappa((t-s)/b_n),$$

$y = [y_{t+H}, \ldots, y_n]'$, $W = [w_{H+1}, \ldots, w_n]'$ and $Z = [z_{H+1}, \ldots, z_n]'$. The Almon-parameterized

[14] One could argue that the Almon-parameterized KLM statistic is unnecessary even when there are only one or two endogenous variables, because one could simply reduce the order of Almon polynomials to make the model just-identified. However, this approach may incur potentially substantial information loss. Using only one or two Almon polynomials implies approximating the impulse response solely by levels and slopes, which cannot capture a typical hump-shaped impulse response. Rather, the Almon-parameterized KLM statistic allows researchers to include higher-order polynomials to approximate more complex impulse response shapes without sacrificing power or interpretability, unlike the AR statistic.

KLM statistic with the unrestricted long-run variance, denoted $K_a(\delta_0)$, is defined similarly by replacing the restricted estimates $\tilde{s}_u^2$ and $\tilde{s}_{uw}$ with the unrestricted versions defined as in (5) and $\hat{s}_{uw} = \frac{1}{n-H}\hat{u}B_n\widehat{W}$, respectively, where $\widehat{u} = M_Z u, \widehat{W} = M_Z W$, and B_n is an $(n-H) \times (n-H)$ matrix with (s,t) entry equal to $\kappa((t-s)/b_n)$.[15]

Asymptotic theory. The asymptotic framework used in this paper is the same as in Barnichon and Mesters (2020) and nonstandard due to the Almon parameterization. In particular, I let the number of lagged structural shocks entering the Almon-parameterized instruments grow with the sample size such that $H/n \to c \in (0,1)$, which introduces nonstationary instruments with explosive variance.

Barnichon and Mesters (2020) showed that AR_a asymptotically follows $\chi^2(3)$ distribution, and Corollary 1 shows the same follows for AR_a^r.

Barnichon and Mesters (2020) Theorem 1. *Let Assumption 1 hold. Under H_0: $\delta = \delta_0$ for $\frac{H}{n} \to c \in (0,1)$ as $n \to \infty$ we have that*

$$AR_a(\delta_0) \Rightarrow \chi^2(3).$$

Corollary 1. *Under the same conditions as above, we have*

$$AR_a^r(\delta_0) \Rightarrow \chi^2(3).$$

Proof. The only difference between AR_a^r and AR_a is the use of $\tilde{s}_u^2$, so it is sufficient to show its consistency $\tilde{s}_u^2 \xrightarrow{p} \omega_{u,0}^2$, where $\omega_{u,0}^2$ is the long-run variance of u_t, which indeed follows by Lemma 5 of Barnichon and Mesters (2020). □

As for K_a^r and K_a, both of their limiting distributions are $\chi^2(m)$ distributions, under the standard asymptotics, as well as the weak-instrument asymptotics à la Staiger and Stock (1997) and also when the instruments are irrelevant.

[15] For K_a^r, one could instead use $\hat{s}_u^2$ and $\hat{s}_{uw}$ for $\tilde{\Pi}(\delta_0)$. The results obtained in this paper remain quantitatively similar under this alternative.

Theorem 1. *Let Assumption 1 hold. Under H_0: $\delta = \delta_0$ for $\frac{H}{n} \to c \in (0,1)$ as $n \to \infty$, we have*

$$K_a^r(\delta_0), K_a(\delta_0) \Rightarrow \chi^2(m)$$

when

(i) the instruments are relevant such that Π is fixed and has full rank

(ii) the instruments are weakly relevant such that $\Pi = \Pi_n = K_n^{-1}C$ with C being fixed and full rank and $K_n = \mathrm{diag}(n, n^2, n^3)$

(iii) the instruments are irrelevant such that $\Pi = 0$.

Proof. See Supplementary Appendix B. □

Although the scaling by K_n in case (ii) may look different from the standard local-to-zero weak-instrument asymptotics of Staiger and Stock (1997), which scales by $\sqrt{n}$, it is effectively equivalent to their framework adjusted for the Almon-parameterized instruments. This can be seen by looking at the signal component of the concentration parameter $\mu := \Pi'(Z'Z)\Pi$, which measures the instrument strength and can be interpreted as an effective sample size.[16] As in the standard weak-instrument asymptotics, it does not diverge but instead converges: $\mu = C'(K_n^{-1}Z'ZK_n^{-1})C \Rightarrow C'\Psi C$, where Ψ is a random matrix defined in Supplementary Appendix B, thereby preserving finite-sample features asymptotically. In other words, since the Almon parameterization inflates the signal ξ_t by taking the weighted partial sums, and the variances of these weighted sums diverge at different speeds, we need to counteract this artificial inflation by scaling with K_n to make the information content comparable to the standard case. More formally, the signal of the Almon-parameterized instrument per observation under the scaling K_n is equivalent to that of a stationary instrument under the standard weak-instrument asymptotics of Staiger and Stock (1997). To see this, note that the signal per observation in the standard weak-instrument asymptotics is $O_p(n^{-1/2})$, because the stationary instrument is $O_p(1)$ and the scaling is $O(n^{-1/2})$. Then, for the Almon-parametrized instruments, we

[16] The concentration parameter itself is defined as $\mu_{conc} := \Omega_{vv,0}^{-1/2}\Pi'(Z'Z)\Pi\Omega_{vv,0}^{-1/2}$ where $\Omega_{vv,0}$ is the long-run variance of v_t.

have $n^{-1/2}\sum_{h=0}^{H}\xi_{t-h} = O_p(1)$, $n^{-3/2}\sum_{h=0}^{H}h\xi_{t-h} = O_p(1)$, and $n^{-5/2}\sum_{h=0}^{H}h^2\xi_{t-h} = O_p(1)$ by the Functional Central Limit Theorem, as shown in the proof of Lemma 4 in Barnichon and Mesters (2020). Therefore, the signal of each term in the scaled Almon-parameterized instruments, $K_n^{-1}z_t$, is $O_p(n^{-1/2})$.

Since Assumption 1 does not impose any conditions on the instrument strength, and Theorem 1 provides the limiting distributions of K_a^r and K_a under both strong and weak-instrument asymptotics, we can construct weak-instrument-robust confidence sets for δ by inverting them using χ^2 critical values.[17] More formally, the confidence sets constructed as $C_{1-\alpha}^{AR} = \left\{\delta_0 : AR(\delta_0) \leq \chi^2_{1-\alpha}(3)\right\}$ for $AR \in \{AR_a^r, AR_a\}$ and $C_{1-\alpha}^{K} = \left\{\delta_0 : K(\delta_0) \leq \chi^2_{1-\alpha}(m)\right\}$ for $K \in \{K_a^r, K_a\}$ ensure asymptotically correct coverage for the significance level α regardless of instrument strength, where $\chi^2_{1-\alpha}(d)$ is the $(1-\alpha)$ quantile of the χ^2 distribution with d degrees of freedom.However, a caveat applies when inverting the KLM statistic, as noted by Kleibergen (2002). Since the KLM statistic is a quadratic form of the score of the concentrated log-likelihood, it achieves zero not only at the maximum but also at the minimum. Hence, Kleibergen (2002) recommends discarding the finite subsets of confidence sets that do not contain the LIML (limited information maximum likelihood) estimate.[18]

Finite-sample size accuracy. Even though $\{AR_a^r, AR_a\}$ and $\{K_a^r, K_a\}$ are asymptotically equivalent, their finite-sample performance is different. Specifically, the use of restricted versus unrestricted long-run variance estimators leads to an important distinction in finite-sample size accuracy.

Inverting AR_a^r and K_a^r with χ^2 critical values is expected to control size accurately in finite samples. As Lazarus et al. (2018) and Vogelsang (2018) demonstrated within a regression framework, imposing a null hypothesis when computing the long-run variance

[17] Since the number of parameters is at most 3 in this paper, inverting these statistics with grid search is computationally demanding but not prohibitive. Furthermore, the computational cost of the statistics using the restricted long-run variance is slightly smaller compared to the unrestricted variance because it skips the computation of the OLS estimate of θ_a. This contrasts with the conventional intuition in GMM estimation, e.g., a continuous-updating estimator, which precludes analytical solutions and increases computational burden.

[18] Kleibergen (2007) also proposes a pre-test to overcome this spurious bahevior.

helps reduce finite-sample size distortion. I confirm that this finding holds for AR_a^r and K_a^r through Monte Carlo simulations in the next section.

On the other hand, inverting AR_a and K_a with χ^2 critical values may induce over-rejection. As is well established in the HAR inference literature (Müller, 2014; Lazarus et al., 2018), standard hypothesis testing based on the unrestricted long-run variance estimator and conventional χ^2 critical values induces size distortions in finite samples. In this case, applying the critical values from fixed-b asymptotics (Kiefer and Vogelsang, 2002b,a, 2005), which provide a higher-order refinement of a conventional first-order small-b asymptotics (Sun, 2014), may help avoid over-rejection.[19] I also confirm these conjectures in the Monte Carlo simulations coming next.

3 Monte Carlo simulations

3.1 Simulation setup

The simulation setup closely follows Barnichon and Mesters (2020) and is motivated by the empirical characteristics of the hybrid Phillips curve in the U.S. I simulate data from

$$\pi_t = \gamma_b \pi_{t-1} + \gamma_f E_t[\pi_{t+1}] + \lambda x_t + e_t \tag{10}$$

$$x_t = \rho_1 x_{t-1} + \rho_2 x_{t-2} + \varepsilon_t + \nu e_t$$

where $e_t = \rho e_{t-1} + \sqrt{1-\rho^2}\zeta_t$ with $\zeta_t \sim N(0,1)$ and $\varepsilon_t \sim N(0, \sigma_i^2)$. I consider two different parameter settings for γ_f and γ_b. The first is $\gamma_f = 0.4, \gamma_b = 0.6$, which is consistent with the common restriction in the empirical literature that $\gamma_b + \gamma_f = 1$. Under the common

[19] Under standard regression and GMM assumptions, Kiefer and Vogelsang (2002b, 2005) showed that, as the sample size goes to infinity with a fixed kernel bandwidth $b \in (0,1]$, the non-parametric long-run variance estimator weakly converges to a random matrix that is proportional to the long-run variance, depends on the bandwidth and the kernel, and is driven by a Brownian bridge. Using Monte Carlo simulations, they demonstrate that the fixed-b asymptotics yield better size properties in finite samples than the small-b asymptotics, where the bandwidth approaches zero. Sun (2014) confirms this finding theoretically by showing that the fixed-b asymptotics provide a higher-order refinement of the conventional first-order small-b asymptotics. For the non-standard fixed-b asymptotic distribution, we can use the standard F-approximation provided by Sun (2014). In applications, Lewis and Mertens (2024), for instance, use the fixed-b critical values of Sun (2014) for the KLM statistic.

restriction, equation (10) can be mapped to the general form (1) by defining

$$y_t = \pi_t - \pi_{t-1}, \quad w_t = (\pi_{t+1} - \pi_{t-1}, x_t)', \quad \delta = (\gamma_f, \lambda)', \quad \text{and } u_t = e_t + \gamma_f(E_t\pi_{t+1} - \pi_{t+1}),$$

which reduces the number of parameters from 3 to 2, making the model overidentified. The second is $\gamma_f = 0.3, \gamma_b = 0.6$, following the setup in Barnichon and Mesters (2020). The remaining parameters follow Barnichon and Mesters (2020): $\lambda = 0.4$, $\rho_1 = 1.2$, $\rho_2 = -0.4$, $\nu = -1$, $\sigma_i \in \{0.1, 0.25, 0.5, 1\}$, and $\rho \in \{0, 0.5\}$. I use $H = 20$ to construct instruments from ε_t. The sample sizes are $T \in \{100, 200\}$ to reflect the sample sizes observed in the empirical analysis. I run $10,000$ simulations and calculate the empirical rejection frequencies.[20]

3.2 Simulation results

Table 1 presents the empirical rejection frequencies of the Almon-parameterized AR statistics (AR_a^r, AR_a), the Almon-parameterized KLM statistics (K_a^r, K_a), and the standard KLM statistic without Almon parameterization (K^r, K), with both the restricted and unrestricted long-run variance estimators, under the null hypothesis. The left columns display results with the common restriction $\gamma_b + \gamma_f = 1$, while the right columns display results without this restriction. Table 1 offers four main findings.[21]

First, both AR_a^r and K_a^r with the χ^2 critical values control size accurately in finite samples across all specifications. The first and the third columns of Table 1 show that both AR_a^r and K_a^r achieve rejection frequencies close to 5% across all specifications under the common restriction $\gamma_b + \gamma_f = 1$. We observe similar patterns when the common restriction is not imposed for AR_a^r, as shown in the first column of the right panel. These results are consistent with findings in the HAR literature, such as Lazarus et al. (2018), who show that using the restricted long-run variance improves finite-sample size accuracy.

Second, using AR_a with $\chi^2(3)$ critical values, as proposed by Barnichon and Mesters

[20] The simulation is conducted using Dynare 4.5.4 (Adjemian, Bastani, Juillard, Mihoubi, Perendia, Ratto, and Villemot, 2011).

[21] The results are robust when the significance levels are 10% and 32%; see Supplementary Appendix C.

			With $\gamma_b + \gamma_f = 1$												Without $\gamma_b + \gamma_f = 1$			
			χ^2						fixed-b						χ^2		fixed-b	
T	σ_i	ρ	AR_a^r	AR_a	K_a^r	K_a	K^r	K	AR_a^r	AR_a	K_a^r	K_a	K^r	K	AR_a^r	AR_a	AR_a^r	AR_a
100	0.10	0.0	5.0	10.4	3.5	8.8	86	91	0.9	4.0	1.2	4.6	73	82	5.1	9.9	0.9	3.8
100	0.25	0.0	4.9	10.4	3.9	8.7	78	82	0.7	3.6	1.3	4.5	64	71	5.5	10.2	1.1	3.8
100	0.50	0.0	5.4	10.1	4.8	8.8	63	62	1.0	3.7	1.8	4.6	48	50	5.2	10.1	1.0	3.7
100	1.00	0.0	6.6	9.5	5.7	8.2	46	34	1.8	3.2	2.6	4.2	33	24	5.1	9.5	0.9	3.2
100	0.10	0.5	5.5	12.4	3.5	10.0	8	21	0.8	5.4	1.1	5.5	2	11	5.5	12.6	0.8	5.1
100	0.25	0.5	5.1	12.1	3.3	9.5	8	19	1.0	5.1	1.0	5.4	2	10	5.4	12.3	0.8	5.1
100	0.50	0.5	5.3	12.1	3.5	9.8	10	20	0.9	4.9	1.0	5.3	4	10	5.3	12.3	0.8	5.2
100	1.00	0.5	6.3	11.6	4.4	9.6	13	21	1.4	4.8	1.8	5.0	7	11	5.3	11.9	0.9	4.9
200	0.10	0.0	5.2	7.6	3.9	6.2	85	87	2.5	4.6	2.3	4.3	79	83	5.0	7.4	2.5	4.4
200	0.25	0.0	5.0	7.5	3.9	6.5	71	72	2.7	4.4	2.5	4.5	64	67	5.0	7.2	2.5	4.3
200	0.50	0.0	4.9	7.2	4.4	6.3	49	46	2.5	4.2	2.6	4.4	42	39	5.1	7.3	2.7	4.3
200	1.00	0.0	5.6	7.1	5.1	6.3	30	20	3.1	4.2	3.5	4.4	24	16	5.2	7.3	2.6	4.3
200	0.10	0.5	5.2	8.5	3.5	6.6	6	11	2.4	5.1	2.0	4.4	3	7	5.2	8.6	2.5	5.2
200	0.25	0.5	5.4	8.7	3.9	6.6	6	11	2.7	5.4	2.3	4.9	3	7	5.3	8.4	2.7	5.3
200	0.50	0.5	5.8	9.1	3.8	6.5	8	12	2.7	5.7	2.5	4.7	5	8	5.4	8.8	2.7	5.3
200	1.00	0.5	5.5	8.2	4.3	6.8	10	12	2.9	5.0	2.8	4.7	7	8	5.5	8.6	2.8	5.4

Note: This table reports the rejection frequencies (%) for $H_0 : \delta = \delta_0$ at a significance level of $\alpha = 0.05$, based on 10,000 simulations. The left panel reports results with the common restriction $\gamma_g + \gamma_f = 1$, while the right panel reports results without the restriction. Each row corresponds to a specific sample size $T \in \{100, 200\}$, standard deviation of the shock $\sigma_i \in \{0.1, 0.25, 0.5, 1.0\}$, and autocorrelation of the disturbance $\rho \in \{0.0, 0.5\}$. Tests are based on the Almon-restricted AR and KLM and the standard KLM statistics with restricted long-run variance (AR_a^r, K_a^r, K^r) or versions with unrestricted long-run variance (AR_a, K_a, K), using critical values either from the χ^2 distribution or from the fixed-b asymptotics of Sun (2014). Under the restriction, the degrees of freedom for the AR and KLM statistics are three and two, respectively. Without the restriction, the degrees of freedom are three.

Table 1: Rejection frequencies (%) for $\alpha = 0.05$

(2020), over-rejects the null hypothesis in finite samples.[22] The second column of the left panel shows that its rejection frequencies exceed the nominal level of 5%, regardless of the parameter values.[23] We observe a similar pattern when the common restriction $\gamma_f + \gamma_b = 1$ is not imposed in the second column of the right panel. These results are also consistent with the well-established findings in the HAR literature that the standard approach of using the unrestricted long-run variance together with χ^2 critical values leads to over-rejection in finite samples. We also observe a similar pattern for K_a.

Third, using the fixed-b critical values helps AR_a and K_a avoid over-rejection. Yet, applying the fixed-b critical values for AR_a^r and K_a^r seems too conservative in this specific example, since their rejection probabilities are well below the nominal level. Hence, I use AR_a and K_a with the fixed-b critical values as a robustness check in the empirical applications.

Lastly, the KLM statistic without Almon parameterization (K, K^r) over-rejects the null hypothesis regardless of whether the restricted or unrestricted long-run variance is used, and regardless of the use of either χ^2 or fixed-b critical values. This result is consistent with Barnichon and Mesters (2020), who find that the AR test without Almon parameterization over-rejects the null due to the many-instrument problem.

[22] This result may seem contradictory to Barnichon and Mesters (2020) because they state that "our proposed Almon-parameterized AR_a test has correct size regardless of the strength of the instruments and the value of H" (p. 2274). The discrepancy arises because Barnichon and Mesters (2020) use a different test statistic, which is a variant of AR_a^r, from the AR_a proposed in their Theorem 1 in their simulation, without explicitly presenting it. However, note that the statistic in their simulation code differs from AR_a^r in several respects. For example, their code relies on MATLAB's `hac` command, which does not include a truncation lag for the QS kernel and therefore, strictly speaking, does not satisfy Assumption 1 (iv) that $b_n = o(n)$.

[23] The over-rejection of AR_a and K_a becomes negligible when the sample size exceeds $T = 500$, which, however, is impractical to achieve because it requires approximately 125 years of data.

4 Does size accuracy matter in practice? A Phillips curve example

4.1 The hybrid New Keynesian Phillips curve

I study the following hybrid Phillips curve, using the same specification as Barnichon and Mesters (2020):

$$\pi_t = \gamma_b \pi_{t-1}^4 + \gamma_f E_t[\pi_{t+1}^4] + \lambda x_t + \varepsilon_t^c, \tag{11}$$

where π_t is annualized quarter-to-quarter inflation, π_{t-1}^4 is average inflation over the past year, x_t is a forcing variable, such as the unemployment gap or output gap, and ε_t^c is an unobserved exogenous cost-push disturbance. I study two versions of the Phillips curve: one based on equation (11) and the other imposing the common restriction $\gamma_b + \gamma_f = 1$, which makes the model overidentified.

4.2 Data and instrument strength

I use the same dataset as Barnichon and Mesters (2020) for both macroeconomic variables and instruments. The inflation rate is computed from changes in PCE excluding food and energy. There are two types of forcing variables: the unemployment gap and the output gap. Both are obtained as detrended series using the HP filter, with the smoothing parameter set to $1,600$. I use two different types of monetary policy shock proxies as instruments. One is the narrative monetary policy shock proxy of Romer and Romer (2004), available from 1969 to 2007. The other is the high-frequency identified monetary policy surprise, hereafter the HFI proxy, computed by summing the changes in three-month-ahead monthly Fed Funds futures and the 10-year yield around Federal Open Market Committee announcements; this proxy is available from 1990 to 2017. All data are quarterly.

Table 2 summarizes the results of the instrument strength test proposed by Lewis and Mertens (2025). The left columns for each forcing variable show the results for the Phillips curve in equation (11), while the right columns show the results under the restriction $\gamma_f + \gamma_b = 1$. In the table, g denotes the test statistic for the null hypothesis that the bias of the 2SLS estimator exceeds $\tau = 10\%$ of the worst-case benchmark, implying that the instruments are weak, while *cv* denotes the critical value at the 5% significance level. The table shows that both the Romer-Romer and HFI proxies are weak across all specifications, as the test statistics are well below the critical values.

	Unemployment gap				Output gap			
	$\gamma_b, \lambda_U, \gamma_f$		$\gamma_b + \gamma_f = 1, \lambda_U$		$\gamma_b, \lambda_Y, \gamma_f$		$\gamma_b + \gamma_f = 1, \lambda_Y$	
	g	cv	g	cv	g	cv	g	cv
RR	9.3	26.0	10.9	26.0	5.4	25.0	5.8	25.0
HFI	4.6	22.9	4.8	22.9	8.3	21.7	8.8	21.7

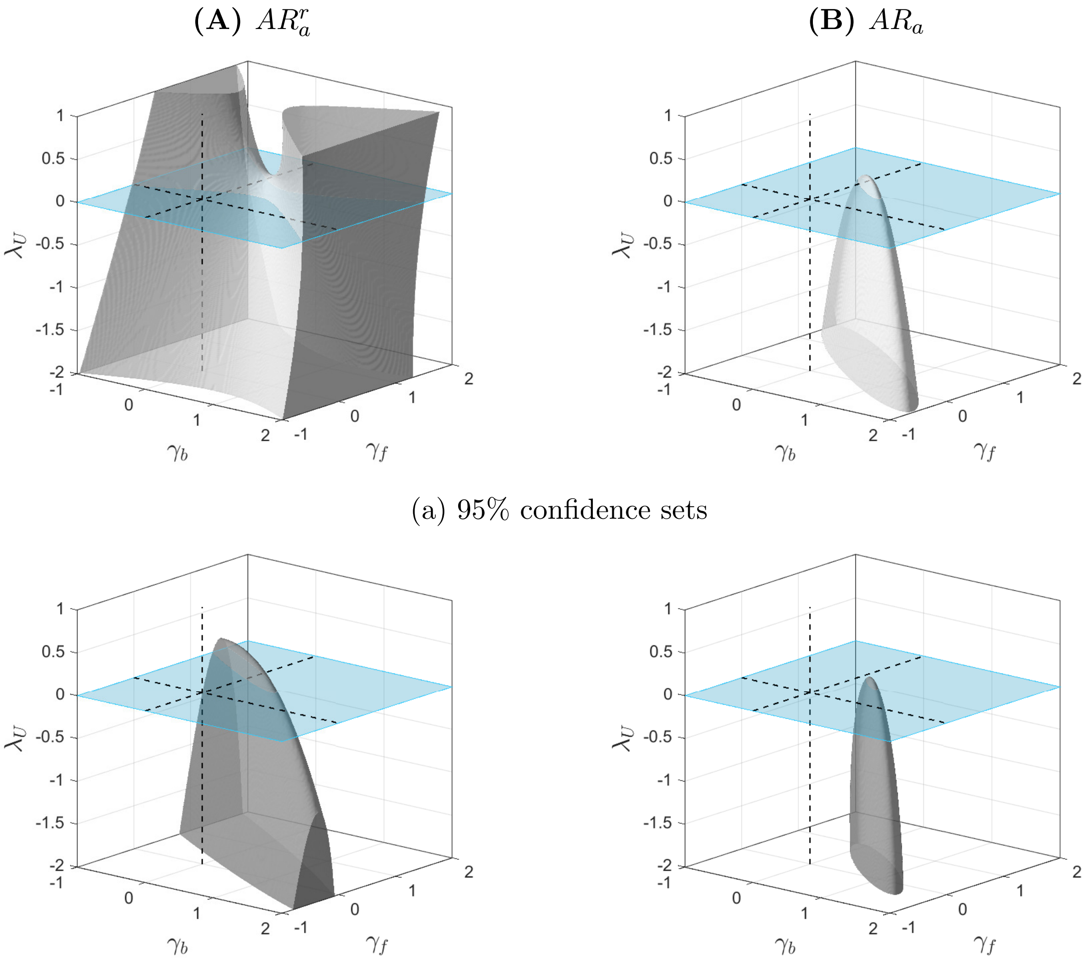
Table 2: Weak instrument test for the Phillips curve

Note: This table presents results of the weak instrument test of Lewis and Mertens (2025). The null hypothesis is that the weak instrument bias exceeds $\tau = 10\%$ of the worst-case benchmark, so that the instruments are weak. g is the test statistic, and the critical values (cv) are for a 5% significance level.

4.3 Results

Phillips curve in 1969–2007 identified by the Romer-Romer proxy. Figure 1 presents the results in 1969–2007 using the Romer-Romer proxy when the forcing variable is the unemployment gap. Panels (a) and (b) plot the 95 and 90% confidence sets over three parameters $\delta = [\lambda_U, \gamma_b, \gamma_f]$, respectively, where the gray areas are the parameter values that are not rejected either by AR_a^r (on the left) or AR_a (on the right) using $\chi^2(3)$ critical values. The plane represents $\lambda_U = 0$, a completely flat Phillips curve. Figure 2 shows the 95, 90, and 68% confidence sets for the two parameters γ_f and λ_U, under the common restriction $\gamma_b + \gamma_f = 1$, using both AR_a^r and AR_a and different critical values for comparison. Here, I use the AR statistic rather than the KLM statistic to ensure direct comparability with Barnichon and Mesters (2020). I confirm that the results are similar when the KLM statistic is used, as shown in Supplementary Appendix D.3. There are three findings from these results.

First, the confidence sets over $\delta = [\lambda_U, \gamma_b, \gamma_f]$ with AR_a^r are substantially larger than



(a) 95% confidence sets

(b) 90% confidence sets

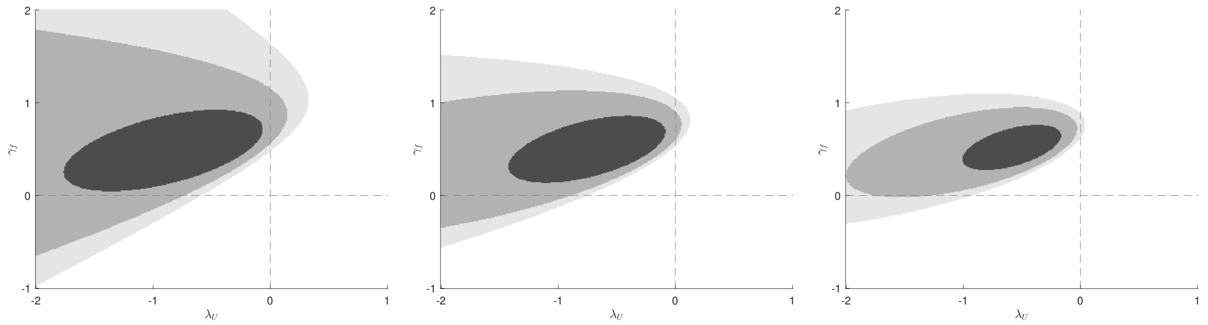
Figure 1: Phillips curve with unemployment gap identified by the Romer-Romer proxy

Note: Panels (a) and (b) plot the 95 and 90% confidence sets over three parameters $\delta = [\lambda_U, \gamma_b, \gamma_f]$, respectively, where the gray areas are the sets that are not rejected either by AR_a^r (on the left) or AR_a (on the right) using $\chi^2(3)$ critical values. The plane represents the area of a completely flat Phillips curve: $\lambda_U = 0$.

those with AR_a, and the quantitative conclusion on the shape of the Phillips curve could change depending on the test statistic. When AR_a^r is used, Panel (a) (A) shows that the confidence sets include both a completely flat and a very steep Phillips curve, for a wider range of γ_f and γ_b; moreover, the confidence sets also cover wider regions for γ_f and γ_b. On the other hand, when AR_a is used, Panel (a) (B) shows that the slope is statistically significant even at the 5% significance level across most of the values of γ_b and γ_f, and γ_b and γ_f are concentrated in small regions around zero to one. This result implies that the seemingly significant results obtained by AR_a are driven mainly by over-rejection, and that the size-correct procedure indicates that we cannot obtain a clear quantitative conclusion about the shape of the Phillips curve from the dataset used in this analysis.

Second, a similar observation follows when we impose the restriction $\gamma_b + \gamma_f = 1$. Specifically, Panel (a) of Figure 2 shows that, with AR_a^r, the 90% confidence set includes both a completely flat and a very steep Phillips curve for a wider range of γ_f. This result is consistent with the findings of Lewis and Mertens (2024), who estimate the same Phillips curve (11) with the common restriction using monthly data and instruments constructed from the main business cycle shock of Angeletos, Collard, and Dellas (2020) and find that the slope coefficient is not significantly different from zero. Furthermore, the 90% confidence set for γ_f includes values even over unity and below zero with AR_a^r for some λ, which implies that the forward-looking coefficients cannot be precisely estimated, either. On the other hand, Panel (b) implies that the result with AR_a may wrongly lead us to conclude that the slope is significantly different from zero at the 10% significance level; and the forward-looking behavior is mostly consistent with theory because γ_f lies within the range of 0 to 1, and is approximately 1 especially when the slope is relatively flat.

Finally, the confidence sets obtained with AR_a under the common restriction (Panel (b) of Figure 2) are much wider than those reported in Figure II (A) of Barnichon and Mesters (2020), which is replicated in Panel (c), even though I use the same dataset and specification. In their figure, the 90% confidence sets exclude the completely flat Phillips curve for all γ_f, and they conclude that the slope is significantly different from zero. Furthermore, they argue that their result “clearly excludes large values for γ_f”. This discrepancy arises



(a) AR_a^r with $\chi^2(3)$ (b) AR_a with $\chi^2(3)$ (c) Barnichon and Mesters

Figure 2: Phillips curve with unemployment gap identified by the Romer-Romer proxy under $\gamma_b + \gamma_f = 1$.

Note: This figure shows the 95% (light gray), 90% (gray), and 68% (black) confidence sets for the two parameters γ_f and λ_U, under the common restriction $\gamma_b + \gamma_f = 1$. Panels (a) and (b) use AR_a^r and AR_a, respectively, with the $\chi^2(3)$ critical values. Panel (c) replicates Figure II (A) of Barnichon and Mesters (2020), which uses AR_a with $\chi^2(2)$ critical values

because they used $\chi^2(2)$ rather than the correct $\chi^2(3)$ for the critical values.[24] Note that the degrees of freedom should be 3 even under the restriction $\gamma_b + \gamma_f = 1$ that reduces the number of parameters from three to two. This is because the degrees of freedom for the limiting χ^2 distribution of the AR test statistic are determined by the number of instruments, which is three under the Almon parameterization, rather than the number of parameters of interest.[25] Therefore, as shown in Panel (a) of Figure 2, their conclusion that the slope of the Phillips curve is significant no longer holds when the misspecification of the degrees of freedom and the finite-sample size distortion of AR_a are corrected.

The results are similar when the forcing variable is the output gap, as shown in

[24] Compared to the results of subset inference without the restriction, Barnichon and Mesters (2020) state that "imposing the common restriction $\gamma_b + \gamma_f = 1$ barely changes ... confidence sets for (λ, γ_f), except that the sets become slightly smaller".However, my results imply that the confidence sets become slightly smaller, not because of the restriction, but because of the use of the $\chi^2(2)$ critical values.

[25] The misconception that the degrees of freedom should be 2 may stem from a misunderstanding of the method as subset inference. Although subset inference is not the focus of this paper, for the sake of argument, the following provides a concise summary of the main idea. Let us partition the parameters δ as $\delta = (\beta', \alpha')'$, and consider testing the null hypothesis $H_0 : \beta = \beta_0$, treating α as nuisance parameters. Theorem 1 of Barnichon and Mesters (2020) shows that, under the null, the subset test statistic, $AR_{a,s}[\beta_0] = \min_{\alpha \in \mathbb{R}^{dim(\alpha)}} AR_a[(\beta_0', \alpha')']$, is asymptotically upper-bounded by a $\chi^2(dim(\beta))$ random variable. Then, some may interpret the analysis imposing the common restriction $\gamma_b + \gamma_f = 1$ as subset inference with $\beta = (\gamma_f, \lambda)'$ and the nuisance parameter $\alpha = \gamma_b = 1 - \gamma_f$, leading them to conclude that $\chi^2(2)$ critical values should be used. This interpretation is not valid, however, because γ_b is no longer a free parameter but is instead fixed given γ_f under the restriction $\gamma_b + \gamma_f = 1$; we cannot minimize over $\alpha = \gamma_b$ to compute the subset test statistic. Therefore, even though we are estimating $\delta = (\gamma_f, \lambda)'$, the test statistic needs to be inverted using $\chi^2(3)$, not $\chi^2(2)$, critical values. In an unreported simulation, available upon request, I compute the rejection frequencies of AR_a with $\chi^2(2)$ critical values under the same simulation setups as Section 3. I find that it rejects the null hypothesis approximately 16% to 21%, depending on the parameter values, when the nominal size is set to 5% for $T = 100$.

Supplementary Appendix D.1. Furthermore, these results are robust when AR_a with fixed-b critical values is used, as shown in Supplementary Appendix F.1.

Phillips curve in 1990–2017 identified by the HFI proxy. Similar findings emerge for the Phillips curve estimated over the 1990–2017 period using the HFI proxy; a complete description of the results is provided in Supplementary Appendix D.2. The confidence sets generated by AR_a^r are substantially wider than those from AR_a. The confidence sets with AR_a^r cover broad regions where parameter values are inconsistent with theoretical predictions, such as a positive λ_U and a negative λ_Y or values of γ_f and γ_b that are negative or exceed unity. Furthermore, the result under the common restriction with AR_a shows larger confidence sets than those in Figure V of Barnichon and Mesters (2020) due to the same reason discussed above.

5 Empirical application: Fiscal multipliers

Having confirmed the effectiveness of AR_a^r and K_a^r in controlling finite-sample size and the practical importance of size accuracy, I now apply these methods to analyze fiscal policy rules and multipliers.

5.1 Model

I begin with a simple fiscal policy rule in which the fiscal instrument (p_t)—either tax revenue (tr_t) or government spending (g_t)—responds systematically to output (y_t) alone:

$$p_t = \psi_{gdp}^p y_t + \varepsilon_t^p \tag{12}$$

where ε_t^p is a fiscal policy shock. Identifying this simple rule is crucial because, as shown by Caldara and Kamps (2017), there exists a one-to-one mapping between ψ_{gdp}^p and the impact fiscal multiplier, given a reduced-form VAR. Specifically, this relationship is given by

$$M_0(\psi^p_{gdp}) = \frac{\sigma_{p,gdp} - \psi^p_{gdp}\sigma^2_{gdp}}{(\psi^p_{gdp}\sigma_{gdp})^2 + \sigma^2_p - 2\psi^p_{gdp}\sigma_{p,gdp}}, \tag{13}$$

where M_0 is the impact fiscal multiplier; and σ^2_{gdp}, σ^2_p, and $\sigma_{p,gdp}$ denote the unconditional variance and covariances of the reduced-form residuals for output and the policy variable in the VAR.

Using this mapping, I adopt a two-step approach to measure the fiscal multiplier. First, I obtain robust confidence sets for ψ^p_{gdp} using K^r_a, since the model is overidentified. Then, I compute the impact multiplier associated with the robust confidence set using (13). This strategy is more robust to weak identification and misspecification than the approach of Caldara and Kamps (2017), which relies on Bayesian inference assuming strong identification and requires the full estimation of a five-variate proxy SVAR.[26] Still, my strategy has limitations relative to their approach; for instance, dynamic fiscal multipliers are not readily available, and the method does not accommodate different sample periods for the data and instruments.

I then extend the analysis to a general fiscal policy rule that allows the fiscal instrument to respond to inflation (π_t) and the real interest rate (r_t) in addition to output:

$$p_t = \psi^p_{gdp} y_t + \psi^p_\pi \pi_t + \psi^p_r r_t + \varepsilon^p_t. \tag{14}$$

Caldara and Kamps (2017) argue that the fiscal multiplier is mostly determined by ψ^p_{gdp} even under the general rule, so estimating (14) can serve as a robustness check. Furthermore, the rule itself is of interest for fiscal analysis, because of its structural interpretation. For instance, the sign and magnitude of ψ^{tr}_π could be related to bracket creep.

[26] The treatment of weak identification in the Bayesian framework has been under development; see Giannone, Lenza, and Primiceri (2025) and references therein. Weak-instrument-robust approaches to the proxy SVAR model have recently been developed, for instance, by Montiel Olea, Stock, and Watson (2021); Jentsch and Lunsford (2022); Angelini, Cavaliere, and Fanelli (2024).

5.2 Data and instrument strength

I use the same dataset as Caldara and Kamps (2017) for both the macroeconomic variables and the instruments. For the five-variate reduced-form VAR used to obtain the mapping (13), I use real federal tax revenue and government spending (government consumption and investment) per capita, both deflated by the GDP deflator; real GDP per capita; consumer price inflation; and the 3-month T-Bill rate. I use the same data for the simple and general fiscal rules. The VAR has four lags. Tax revenue, spending, and GDP are detrended using the log-deviation from a deterministic linear time trend. The sample period is from 1950 to 2007.[27]

I use two types of instruments to identify the simple and general fiscal policy rules. To identify the simple rule, I use the utilization-adjusted TFP proxy of Fernald (2014), following Caldara and Kamps (2017). To identify the general rule, I use either the TFP proxy or the oil-price-shock proxy of Hamilton (2003), which is computed as a non-linear transformation of the quarterly change in the crude petroleum producer price index.

Although Caldara and Kamps (2017) use the Romer-Romer monetary policy shock proxy, which is available only from 1969 to 2007, in addition to the TFP and oil proxies, I do not use it in my analysis.[28] There are two reasons for this. First, while the monetary policy proxy is required in the framework of Caldara and Kamps (2017) to identify the general rule, it is unnecessary for my approach. In particular, their framework requires monetary policy proxies and a short-run restriction to fully identify the five-variate SVAR, whereas my approach requires only a single shock proxy. The Almon parameterization generates three instruments from a single proxy, which is sufficient to identify the general rules. Second, my method does not accommodate different sample periods for instruments, whereas the proxy SVAR of Caldara and Kamps (2017) does.

[27] The framework proposed in this paper could accommodate state-dependent multipliers by expanding the sample periods to include multiple recessions (Auerbach and Gorodnichenko, 2012; Owyang, Ramey, and Zubairy, 2013; Ramey and Zubairy, 2018; Inoue, Rossi, and Wang, 2024), the effective lower bound period (Miyamoto, Nguyen, and Sergeyev, 2018), and possibly different regimes of monetary-fiscal interaction (Cloyne, Jorda, and Taylor, 2020). I do not take this direction, however, due to the limited strength of the current instruments, which makes it harder to statistically distinguish state-dependent multipliers even if they exist. Exploring this direction is left for future research.

[28] Angelini et al. (2023) revisit the analysis by Caldara and Kamps (2017), and they do not use the Romer-Romer proxy in their main analysis "to circumvent possible parameter instabilities".

Table 3 presents the result of the weak-instrument test of Lewis and Mertens (2025). For the simple fiscal policy rules, the TFP proxy is not strong enough, as the test statistic is well below the critical values. For the general fiscal policy rules, both the TFP and oil proxies are similarly weak.

	Tax				Spending			
	Simple		General		Simple		General	
	g	cv	g	cv	g	cv	g	cv
TFP	3.2	11.4	0.01	23.3	3.2	11.3	0.01	23.5
Oil	-	-	2.5	35.1	-	-	2.5	35.3

Table 3: Weak instrument test for the fiscal policy rule

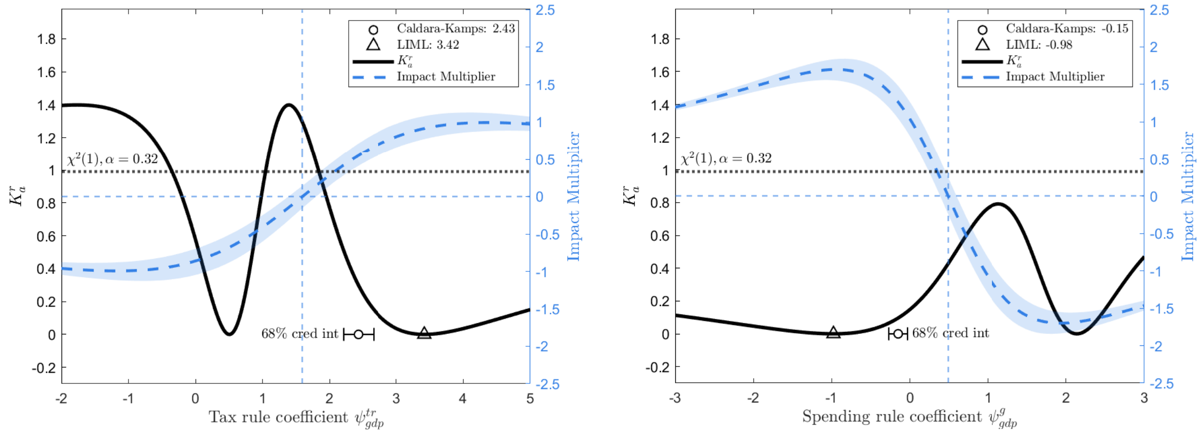
Note: This table presents results of the weak instrument test of Lewis and Mertens (2025). The null hypothesis is that the weak instrument bias exceeds $\tau = 10\%$ of the worst-case benchmark, so that the instruments are weak. g is the test statistic, and the critical values (cv) are for a 5% significance level.

5.3 Results

The simple fiscal rules and multipliers. Figure 3 presents the K_a^r statistic (solid line, left axis) and the impact multiplier (dashed line, right axis) calculated by equation (13) as a function of the systematic response of fiscal policy to output for the simple tax (ψ_{gdp}^{tr}) and spending (ψ_{gdp}^{g}) rules. The dashed vertical line indicates the value of the systematic response for which the impact multiplier is zero. The shaded area around the impact multiplier is the pointwise confidence interval obtained from $10,000$ bootstrap samples, while the dotted horizontal line represents the critical value for K_a^r based on the $\chi^2(1)$ distribution. Following Caldara and Kamps (2017), I adopt a 32% significance level for the analysis. There are five findings from this analysis.[29]

First, the negative and slightly positive tax coefficients are rejected, implying that the tax multiplier is significantly positive. Panel (a) of Figure 3 shows that values of ψ_{gdp}^{tr} ranging from -2 to approximately 2 are rejected at the 32% significance level, while the coefficients ranging from approximately 2 to 5 are not rejected. Note that the low value of K_a^r around $\psi_{gdp}^{tr} = 0.5$ is caused by the *minimum* of the concentrated log-likelihood,

[29] These results are robust when K_a with the fixed-b critical values is used, as shown in Supplementary Appendix F.2.



(a) Simple tax rule

(b) Simple spending rule

Figure 3: K_a^r statistic (left axis), and the impact fiscal multiplier (right axis)

Note: This figure presents the K_a^r statistic (solid line) and the impact multiplier calculated using equation (13) (dashed line) as a function of the systematic response of fiscal policy to output. The triangle represents the limited-information maximum likelihood estimate. The shaded area around the impact multiplier is the pointwise confidence interval obtained from 10,000 bootstrap samples, while the dotted horizontal line indicates the critical value for K_a^r, both at the 32% significance level. The circle and error bars represent the point estimate and the 68% credible interval from Caldara and Kamps (2017). The dashed vertical line indicates the value of ψ_{gdp}^{p}, $p \in \{tr, g\}$, where the impact multiplier is zero. Panels (a) and (b) display the results for the simple tax and spending rules, respectively.

as this area does not include the LIML estimate; hence, it should be discarded from the confidence set. Applying the mapping (13), this result implies that the negative multiplier is rejected, while the positive multiplier is not rejected. This result, to my knowledge, provides the first statistically significant evidence regarding the tax multiplier that accounts for weak-instrument issues. Furthermore, K_a^r achieves its minimum and is mostly flat around $\psi_{gdp}^{tr} = 3.4$, and these values of ψ_{gdp}^{tr} imply a multiplier of approximately one.

Second, the spending coefficient and multiplier cannot be precisely estimated. Panel (b) of Figure 3 shows that K_a^r does not exceed the critical value anywhere for $\psi_{gdp}^{g} \in [-3, 3]$. Still, the LIML estimate is -0.98 and the K_a^r statistic is almost flat for $\psi_{gdp}^{g} \in (-2, 0)$, which corresponds to an impact multiplier between one and two. Hence, although the data cannot distinguish among a wide range of alternatives, the point estimate and shape of K_a^r suggest a spending multiplier moderately above one.

Third, the point estimates and 68% credible intervals of Caldara and Kamps (2017) align with the results presented here. Their point estimate and the credible interval fall within the robust confidence sets for both tax and spending coefficients. Moreover, their point estimates are located near the minima of the K_a^r statistic. However, their credible intervals are substantially narrower than the robust confidence sets.[30] This difference likely stems from the different assumptions about the strength of identification: Caldara and Kamps (2017) assume strong identification, whereas the confidence sets derived in this paper are robust to weak instruments.

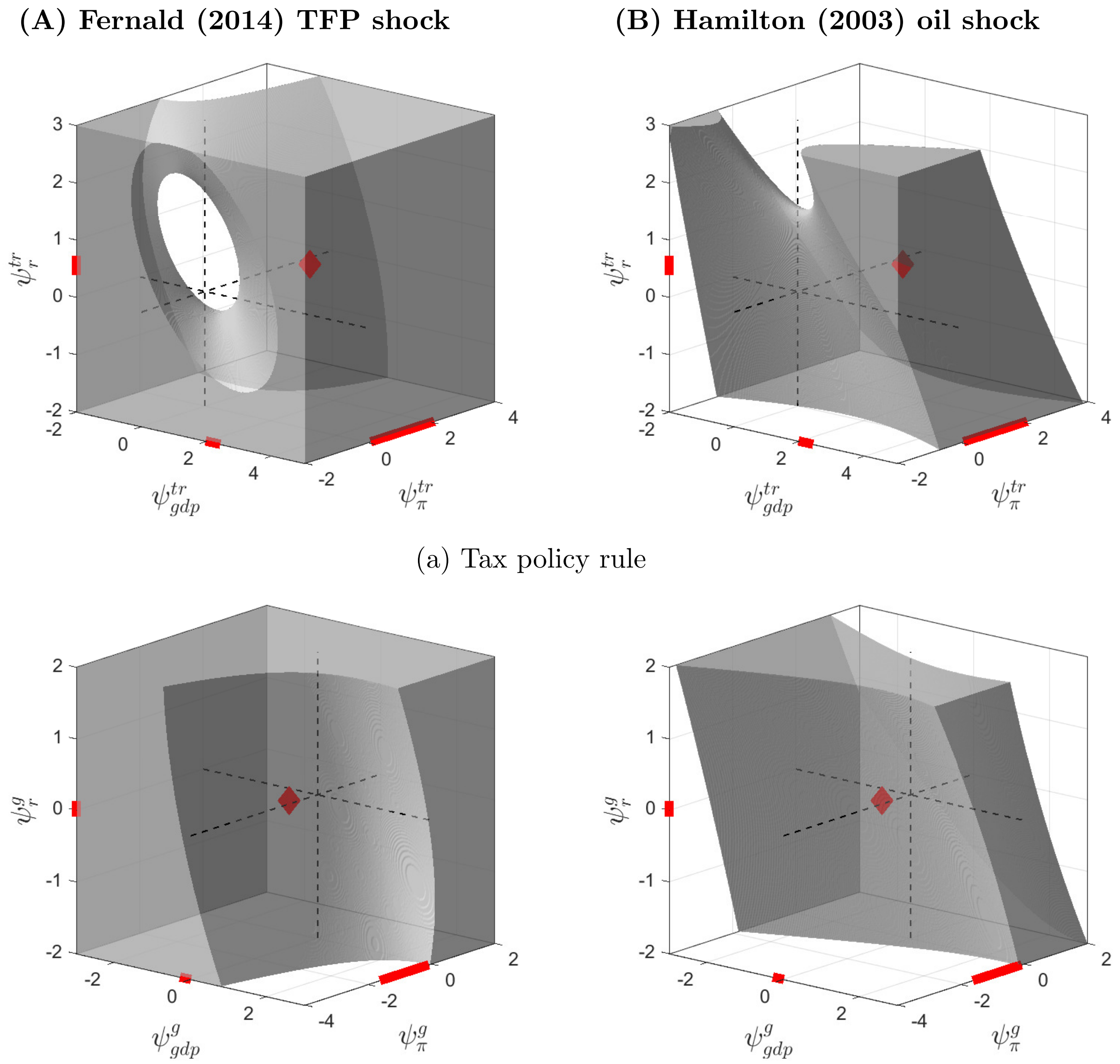
Fourth, these results have implications for the SVAR approach with short-run restrictions. For instance, the popular approach of Blanchard and Perotti (2002) restricts $\psi_{gdp}^{tr} = 2.08$ and $\psi_{gdp}^{g} = 0$ for identification. For the tax rule, although this restriction is not marginally rejected at the 32% significance level, Figure 3 implies that pinpointing a specific value is difficult because the K_a^r statistic is relatively flat for the values of ψ_{gdp}^{tr} larger than 2. For the spending rule, this restriction also does not contradict the results, but a negative ψ_{gdp}^{g} is equally consistent with the data, since the K_a^r statistic is almost flat in this region. Overall, drawing quantitative conclusions about the tax policy rule is difficult when accounting for weak instruments, and even more so for the spending rule. Researchers implementing short-run restrictions should therefore check robustness across different assumptions on ψ_{gdp}^{tr} and ψ_{gdp}^{g}.

Lastly, similar to the results of the Phillips curve, the confidence sets become narrower when K_a is used with χ^2 critical values, as shown in Supplementary Appendix E. While the change is not large enough to affect the quantitative conclusion for this case, this result implies that using the size-accurate procedure is essential to avoid spuriously narrow confidence sets.

General fiscal rule. Figure 4 presents the 68% confidence sets for the general tax and spending fiscal rules identified by either the TFP proxy of Fernald (2014) or the oil proxy of Hamilton (2003). The point estimate of Caldara and Kamps (2017) is represented by the diamond (triangle) in the figure if it is (is not) included in the confidence sets, and their credible intervals are depicted on the axes. The results confirm the findings from

[30] Note that the Bayesian credible intervals and the frequentist confidence sets are conceptually different, but numerically they may coincide asymptotically by the Bernstein-von Mises theorem.

the simple fiscal policy rules.[31]


(A) Fernald (2014) TFP shock

(B) Hamilton (2003) oil shock

(a) Tax policy rule

(b) Spending policy rule

Figure 4: 68% confidence sets for the general fiscal policy rules

The gray area represents the 68% confidence sets calculated by AR_a^r with $\chi^2(3)$ critical values. The red lines on the axes show the 68% credible intervals of Caldara and Kamps (2017). The point estimate of Caldara and Kamps (2017) is shown as a red diamond (triangle) if it is (not) included in the confidence set. Panels (a) and (b) correspond to the general tax and spending policy rules, respectively, and columns (A) and (B) show the results identified using the Fernald (2014) TFP proxy and the Hamilton (2003) oil proxy, respectively.

First, the results of Caldara and Kamps (2017) are consistent with Figure 4, especially for the sign of the systematic response to output, the key determinant of the fiscal multiplier. The point estimates of Caldara and Kamps (2017) are all included in the

[31] The results are quantitatively similar when AR_a with fixed-b critical values is used, and the point estimate of Caldara and Kamps (2017) is included in the robust confidence sets, except for the tax policy rule identified by the oil proxy. See Supplementary Appendix F.2 for details.

confidence sets regardless of the specification, and their credible sets are also mostly contained in the confidence set. In particular, the signs of ψ^{tr}_{gdp} and ψ^{g}_{gdp} are consistent with Caldara and Kamps (2017). For instance, in Panel (a) (A), which shows the general tax rule identified by the TFP proxy, the accepted areas concentrate where ψ^{tr}_{gdp} is positive. Similarly, in Panel (a) (B), the accepted area is slightly larger for positive ψ^{tr}_{gdp} than for negative, especially when ψ^{tr}_{r} is positive. We observe similar patterns for the spending policy rule, where a negative ψ^{g}_{gdp} is generally supported.

Second, the robust confidence sets are too wide to draw quantitative conclusions about general fiscal policy rules. When the TFP proxy is used for identification, the accepted region covers most of the area for both the tax and spending rules. The accepted region with the oil proxy is slightly smaller, but ψ^{tr}_{r} could take both positive and negative signs for most values of ψ^{tr}_{π} and ψ^{tr}_{gdp}, for instance. A similar observation follows for the spending rule.

Third, the confidence sets become smaller when AR_a is used with the χ^2 critical values, as shown in Supplementary Appendix E. While using AR_a does not change the quantitative conclusion in this case, similar to the simple fiscal rule, this result confirms the importance of using a size-accurate procedure to avoid overly narrow confidence sets.

6 Conclusion

This paper proposes an Almon-parameterized KLM statistic for overidentified models and a method to accurately control finite-sample size for both the Almon-parameterized AR and KLM statistics. The empirical importance of size accuracy is highlighted by the Phillips curve application: existing methods, which suffer from size distortion, yield statistically significant conclusions about the slope of the Phillips curve, whereas size-accurate methods challenge these findings. Furthermore, by applying the method to fiscal policy rules, I find that the tax multiplier is significantly positive, although it remains difficult to pinpoint the exact magnitudes of the systematic components of tax and spending policy rules due to large uncertainty. These results imply that, as noted by Mavroeidis et al.

(2014) in the context of the Phillips curve,[32] current data and identification strategies still face limitations in providing precise quantitative conclusions when accounting for weak-instrument issues. Consequently, new identification approaches, estimation methods, and datasets are required to reach an empirical consensus on the Phillips curve and fiscal rules.

In future work, comparing the Almon-parameterized KLM test statistic with other popular approaches for overidentified models, such as the conditional likelihood ratio test of Moreira (2003), would provide further guidance for applied studies. Developing criteria for selecting the optimal lag length H, the polynomial type, and the polynomial order is also important.[33] Finally, extending the current Almon-parameterized KLM statistic to subset inference when the nuisance parameters are weakly identified is another fruitful direction.[34]

References

Adjemian, S., H. Bastani, M. Juillard, F. Mihoubi, G. Perendia, M. Ratto, and S. Villemot (2011): "Dynare: Reference Manual, Version 4," Dynare Working Papers 1, CEPREMAP.

Almon, S. (1965): "The distributed lag between capital appropriations and expenditures," *Econometrica: Journal of the Econometric Society*, 178–196.

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Angeletos, G.-M., F. Collard, and H. Dellas (2020): "Business-Cycle Anatomy," *American Economic Review*, 110, 3030–3070.

[32] Specifically, Mavroeidis et al. (2014) state that "the literature has reached a limit on how much can be learned about the New Keynesian Phillips curve from aggregate macroeconomic time series. New identification approaches and new datasets are needed to reach an empirical consensus."

[33] Lewis and Pal Mustafi (2026) recently proposed one direction on this issue.

[34] Recently, Londschien and Bühlmann (2026) proposed a subvector KLM statistic under homoskedastic errors, building on Guggenberger, Kleibergen, Mavroeidis, and Chen (2012). Their approach requires either a specific technical condition to hold or the number of instruments to grow to infinity with the sample size.

Angelini, G., G. Caggiano, E. Castelnuovo, and L. Fanelli (2023): "Are Fiscal Multipliers Estimated with Proxy-SVARs Robust?" *Oxford Bulletin of Economics and Statistics*, 85, 95–122.

Angelini, G., G. Cavaliere, and L. Fanelli (2024): "An Identification and Testing Strategy for Proxy-SVARs with Weak Proxies," *Journal of Econometrics*, 238, 105604.

Ascari, G., Q. Haque, L. M. Magnusson, and S. Mavroeidis (2024): "Empirical Evidence on the Euler Equation for Investment in the US," *Journal of Applied Econometrics*, 39, 543–563.

Ascari, G., L. M. Magnusson, and S. Mavroeidis (2021): "Empirical Evidence on the Euler Equation for Consumption in the US," *Journal of Monetary Economics*, 117, 129–152.

Auerbach, A. J. and Y. Gorodnichenko (2012): "Measuring the Output Responses to Fiscal Policy," *American Economic Journal: Economic Policy*, 4, 1–27.

Barnichon, R. and G. Mesters (2020): "Identifying Modern Macro Equations with Old Shocks," *The Quarterly Journal of Economics*, 135, 2255–2298.

——— (2023): "A sufficient statistics approach for macro policy," *American Economic Review*, 113, 2809–2845.

Blanchard, O. and R. Perotti (2002): "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output," *The Quarterly Journal of Economics*, 117, 1329–1368.

Caldara, D. and C. Kamps (2017): "The Analytics of SVARs: A Unified Framework to Measure Fiscal Multipliers," *The Review of Economic Studies*, 84, 1015–1040.

Carlevaro, E., Q. Haque, and L. Magnusson (2025): "Empirical Evidence on the US Monetary-Fiscal Policy Mix," Tech. rep., University of Adelaide.

Chernozhukov, V., C. Hansen, and M. Jansson (2009): "Admissible Invariant Similar Tests for Instrumental Variables Regression," *Econometric Theory*, 25, 806–818.

Cloyne, J. S., O. Jorda, and A. M. Taylor (2020): "Decomposing the Fiscal Multiplier," NBER Working Paper 26939, National Bureau of Economic Research.

Davidson, J. (2021): *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford University Press.

De Jong, R. M. and J. Davidson (2000): "Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices," *Econometrica*, 68, 407–423.

Dovì, M.-S. (2025): "Inference with High-Dimensional Weak Instruments and the New Keynesian Phillips Curve," *Journal of Business & Economic Statistics*, 1–13, advance online publication.

Fernald, J. (2014): "A Quarterly, Utilization-Adjusted Series on Total Factor Productivity," Working Paper 2012-19, Federal Reserve Bank of San Francisco.

Furlanetto, F. and A. Lepetit (2025): "The Slope of the Phillips Curve," in *Research Handbook on Inflation*, Edward Elgar Publishing, 167–185.

Galí, J. and M. Gertler (1999): "Inflation Dynamics: A Structural Econometric Analysis," *Journal of Monetary Economics*, 44, 195–222.

Giannone, D., M. Lenza, and G. E. Primiceri (2025): "Bayesian Inference in IV Regressions," Discussion Paper DP19920, CEPR.

Gorodnichenko, Y. and B. Lee (2020): "Forecast Error Variance Decompositions with Local Projections," *Journal of Business & Economic Statistics*, 38, 921–933.

Gregory, A. W., J. McNeil, and G. W. Smith (2024): "US Fiscal Policy Shocks: Proxy-SVAR Overidentification via GMM," *Journal of Applied Econometrics*, 39, 607–619.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012): "On the asymptotic sizes of subset Anderson–Rubin and Lagrange multiplier tests in linear instrumental variables regression," *Econometrica*, 80, 2649–2666.

Gürkaynak, R. S., B. Sack, and E. T. Swanson (2005): "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements," *International Journal of Central Banking*, 1, 55–93.

Hamilton, J. D. (2003): "What Is an Oil Shock?" *Journal of Econometrics*, 113, 363–398.

Hansen, L. P. and K. J. Singleton (1982): "Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica: Journal of the Econometric Society*, 1269–1286.

Haque, Q. and L. M. Magnusson (2023): "Identification Robust Empirical Evidence on the Open Economy IS-Curve," *Oxford Bulletin of Economics and Statistics*, 85, 345–372.

Hebous, S. and T. Zimmermann (2018): "Revisiting the Narrative Approach of Estimating Tax Multipliers," *The Scandinavian Journal of Economics*, 120, 428–439.

Inoue, A., B. Rossi, and Y. Wang (2024): "Local Projections in Unstable Environments," *Journal of Econometrics*, 244, 105726.

——— (2025): "Has the Phillips Curve Flattened?" *Econometric Theory*, 1–46, advance online publication.

Jansson, M. (2004): "The Error in Rejection Probability of Simple Autocorrelation Robust Tests," *Econometrica*, 72, 937–946.

Jentsch, C. and K. G. Lunsford (2022): "Asymptotically Valid Bootstrap Inference for Proxy SVARs," *Journal of Business & Economic Statistics*, 40, 1876–1891.

Keweloh, S. A., M. Klein, and J. Prüser (2025): "Estimating Fiscal Multipliers by Combining Statistical Identification with Potentially Endogenous Proxies," *The Econometrics Journal*, utaf027, advance online publication.

Kiefer, N. M. and T. J. Vogelsang (2002a): "Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel Without Truncation," *Econometrica*, 70, 2093–2095.

——— (2002b): "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size," *Econometric Theory*, 18, 1350–1366.

——— (2005): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, 1130–1164.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

——— (2007): "Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics," *Journal of Econometrics*, 139, 181–216.

Kleibergen, F. and S. Mavroeidis (2009): "Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve," *Journal of Business & Economic Statistics*, 27, 293–311.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018): "HAR Inference: Recommendations for Practice," *Journal of Business & Economic Statistics*, 36, 541–559.

Lewis, D. J. and K. Mertens (2024): "Dynamic Identification Using System Projections and Instrumental Variables," Working paper, Federal Reserve Bank of Dallas.

——— (2025): "A Robust Test for Weak Instruments for 2SLS with Multiple Endogenous Regressors," *Review of Economic Studies*, rdaf103.

Lewis, D. J. and U. Pal Mustafi (2026): "Inference for Macroeconomic Policy Counterfactuals," Working paper.

Londschien, M. and P. Bühlmann (2026): "Weak-instrument-robust subvector inference in instrumental variables regression: A subvector Lagrange multiplier test and properties of subvector Anderson-Rubin confidence sets," .

Magnusson, L. M. and S. Mavroeidis (2014): "Identification Using Stability Restrictions," *Econometrica*, 82, 1799–1851.

Mavroeidis, S. (2010): "Monetary Policy Rules and Macroeconomic Stability: Some New Evidence," *American Economic Review*, 100, 491–503.

Mavroeidis, S., M. Plagborg-Møller, and J. H. Stock (2014): "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve," *Journal of Economic Literature*, 52, 124–188.

McKay, A. and C. K. Wolf (2023): "What can time-series regressions tell us about policy counterfactuals?" *Econometrica*, 91, 1695–1725.

Mertens, K. and J. L. Montiel Olea (2018): "Marginal Tax Rates and Income: New Time Series Evidence," *The Quarterly Journal of Economics*, 133, 1803–1884.

Mertens, K. and M. O. Ravn (2013): "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States," *American Economic Review*, 103, 1212–1247.

——— (2014): "A Reconciliation of SVAR and Narrative Estimates of Tax Multipliers," *Journal of Monetary Economics*, 68, S1–S19.

Mikusheva, A. (2021): "Many weak instruments in time series econometrics," in *World Congress of Econometric Society. MIT.*

Miyamoto, W., T. L. Nguyen, and D. Sergeyev (2018): "Government Spending Multipliers Under the Zero Lower Bound: Evidence from Japan," *American Economic Journal: Macroeconomics*, 10, 247–277.

Montiel Olea, J. L., J. H. Stock, and M. W. Watson (2021): "Inference in Structural Vector Autoregressions Identified with an External Instrument," *Journal of Econometrics*, 225, 74–87.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

Müller, U. K. (2014): "HAC Corrections for Strongly Autocorrelated Time Series," *Journal of Business & Economic Statistics*, 32, 311–322.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Owyang, M. T., V. A. Ramey, and S. Zubairy (2013): "Are Government Spending Multipliers Greater During Periods of Slack? Evidence from Twentieth-Century Historical Data," *American Economic Review*, 103, 129–134.

Plagborg-Møller, M. and C. K. Wolf (2022): "Instrumental Variable Identification of Dynamic Variance Decompositions," *Journal of Political Economy*, 130, 2164–2202.

Ramey, V. A. (2016): "Macroeconomic Shocks and Their Propagation," in *Handbook of Macroeconomics*, Elsevier, vol. 2, 71–162.

——— (2019): "Ten Years After the Financial Crisis: What Have We Learned from the Renaissance in Fiscal Research?" *Journal of Economic Perspectives*, 33, 89–114.

Ramey, V. A. and S. Zubairy (2018): "Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data," *Journal of Political Economy*, 126, 850–901.

Romer, C. D. and D. H. Romer (2004): "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 94, 1055–1084.

Smets, F. and R. Wouters (2007): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97, 586–606.

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

Sun, Y. (2014): "Let's Fix It: Fixed-b Asymptotics Versus Small-b Asymptotics in Heteroskedasticity and Autocorrelation Robust Inference," *Journal of Econometrics*, 178, 659–677.

Sun, Y., P. C. B. Phillips, and S. Jin (2008): "Optimal Bandwidth Selection in Heteroskedasticity–Autocorrelation Robust Testing," *Econometrica*, 76, 175–194.

Swanson, E. T. (2021): "Measuring the effects of federal reserve forward guidance and asset purchases on financial markets," *Journal of Monetary Economics*, 118, 32–53.

Varela, M. C. and A. P. F. Ribeiro (2025): "Empirical Literature on Fiscal Multipliers: A Bibliometric Approach, 2002–2023," *Journal of Economic Surveys*, advance online publication.

Vogelsang, T. J. (2018): "Comment on "HAR Inference: Recommendations for Practice"," *Journal of Business & Economic Statistics*, 36, 569–573.

Yogo, M. (2004): "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak," *Review of Economics and Statistics*, 86, 797–810.

Supplementary Appendix (Not for publication)

A Assumptions

Following Barnichon and Mesters (2020), I assume the following conditions for the data-generating process and the kernel.

Assumption 1. *The observations $\{y_t, w_t, z_t\}$ are generated by the linear IV model*

$$y_t = w_t'\delta + u_t \Leftrightarrow y = W\delta + u$$

$$w_t = \Pi' z_t + v_t \Leftrightarrow W = Z\Pi + V$$

and

$$z_t = \left(\sum_{h=0}^{H} \xi_{t-h}, \sum_{h=0}^{H} h\xi_{t-h}, \sum_{h=0}^{H} h^2 \xi_{t-h} \right)'$$

for $t = H+1, \ldots, n$, where δ is $m \times 1$, Π is $k \times m$ with $m \leq k = 3$, and $\eta_t = (\xi_t, u_t, v_t')$. Furthermore, I assume that

(i) for all t and s, we have (a) $E[\eta_t] = 0$, (b) $E[u_t \xi_s] = 0$, and (c) $E[v_t \xi_s] = 0$,

(ii) for some $r > 2$ and finite constant Δ, we have $\sup_t \|\eta_t\|_{2r} \leq \Delta$,

(iii) η_t is L_2-NED (near-epoch dependence)[1] of size $-\frac{r-1}{r-2}$ with $d_t = 1$ on V_t, where $\{V_t\}$ is an α-mixing process of size $-\frac{r}{r-2}$,

(iv) for integers $p, q \geq 0$ we have uniformly in n and H, with $H < n$, that

$$\omega^2_{\xi,p,n,H} = \mathrm{Var}\left[\sum_{t=H+1}^{n} t^p \xi_t \right] = \omega^2_{\xi,p}(n-H)^{2p+1} + o((n-H)^{2p+1})$$

$$\Omega_{uv,q,n,H} = \mathrm{Var}\left[\sum_{t=H+1}^{n} t^q (u_t, v_t')' \right] = \Omega_{uv,q}(n-H)^{2q+1} + o((n-H)^{2q+1}),$$

[1] For the definition of an NED process, see Davidson (2021, Chap. 18).

where $\Omega_{uv,q} = \begin{bmatrix} \omega_{u,q}^2 & \omega_{uv,q} \\ \omega_{vu,q} & \Omega_{vv,q} \end{bmatrix}$ *with finite* $\omega_{\xi,p}^2 > 0$ *and positive definite* $\Omega_{uv,q}$*, and* $\omega_{u,0}^2$ *denotes the long-run variance of* u_t *and* $\omega_{uv,0}$ *denotes the long-run covariance of* u_t *and* v_t*.*

(v) $b_n = o(n)$ *and* $\kappa(\cdot) \in \mathcal{K}$ *where*

$$\mathcal{K} = \left\{ \begin{array}{l} \kappa(\cdot) : \mathbb{R} \to [-1, 1], \kappa(0) = 1, \\ \kappa(x) = \kappa(-x)\ \forall x \in \mathbb{R}, \int_{-\infty}^{\infty} |\kappa(x)| dx < \infty, \\ \int_{-\infty}^{\infty} \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa(x) e^{ivx} dx \right| dv < \infty, \\ \kappa(\cdot) \textit{ is continuous at 0 and at all except for a finite number of points.} \end{array} \right\}$$

Assumption (i) is a standard normalization ensuring that innovations and shock proxies are mean zero and assumes contemporaneous, lead, and lag exogeneity of the shock proxy. Assumptions (ii) and (iii) impose mild restrictions on the dependence and heterogeneity of the innovations and shock proxies and allow heteroskedasticity and serial correlation. Assumption (iv) defines the asymptotic behavior of the long-run variances of η_t. Aside from the inclusion of the polynomial weights t^p and t^q arising from the Almon-parameterization of the instruments z_t, this condition is standard. Assumption (v) is a standard assumption in the HAR literature, specifically following De Jong and Davidson (2000).

B Proof

This section presents the proof of Theorem 1.

B.1 Intermediate results

We need Lemmas A, B, C, and D for completing the proof of Theorem 1. While some of these lemmas build on Lemmas 1 to 5 of Barnichon and Mesters (2020), which are prepared for the just-identified case ($k = m = 3$), they straightforwardly extend to the overidentified cases ($m < k$) by dropping the unnecessary elements of V and related objects, and by adjusting the sizes of vectors and matrices accordingly. Furthermore, the lemmas and the proof straightforwardly extend to cases with the higher-order Almon polynomials and with $m > 3$, by adjusting the matrix dimensions and other related objects, such as the scaling matrix K_n, $D_\xi(a)$, and Ξ_V. For example, one would append n^3 to the diagonal of K_n and calculate $D_{\xi,4}(a)$ in $D_\xi(a)$ when including a third-order polynomial. Additionally, I restate Lemmas 4 and 5 of Barnichon and Mesters (2020), as I use them in the proof.

Lemma A. *Let* $\Psi := \int_c^1 D_\xi(a)D'_\xi(a)da$ *where* $D_\xi(a) = (D_{\xi,1}(a), D_{\xi,2}(a), D_{\xi,3}(a))'$ *with elements*

$$\begin{aligned}
&D_{\xi,1}(a) = G_{\xi,0}(a) - G_{\xi,0}(a-c)\\
&D_{\xi,2}(a) = aG_{\xi,0}(a) - aG_{\xi,0}(a-c) - G_{\xi,1}(a) + G_{\xi,1}(a-c)\\
&D_{\xi,3}(a) = a^2G_{\xi,0}(a) - a^2G_{\xi,0}(a-c) - 2aG_{\xi,1}(a) + 2aG_{\xi,1}(a-c) + G_{\xi,2}(a) - G_{\xi,2}(a-c)
\end{aligned}$$

for a constant $c \in (0,1)$*, where the scalar process* $G_{\xi,p}(a)$ *is a Gaussian process with almost surely continuous sample paths, independent increments and variances*

$$E[G_{\xi,p}(a)^2] = a^{2p+1}\omega^2_{\xi,p},$$

for integers $p \geq 0$*. Also, let* $K_n = \operatorname{diag}(n, n^2, n^3)$*. Then, given Assumption 1, under the null* $H_0 : \delta = \delta_0$ *when* $n \to \infty$ *with* $H/n \to c \in (0,1)$*, we have that*

$$\left[K_n^{-1}Z'(y - W\delta_0), K_n^{-1}Z'\left((W - Z\Pi) - (y - W\delta_0)\tfrac{\omega_{uv,0}}{\omega^2_{u,0}}\right)\right] \Rightarrow [\Xi_u, \Xi_{V\cdot u}]$$

where

$$\operatorname{vec}(\Xi_u, \Xi_{V\cdot u})|_{D_\xi} \sim N\left(0, \begin{bmatrix}\omega^2_{u,0} & 0\\ 0 & \Omega_{V\cdot u}\end{bmatrix} \otimes \Psi\right)$$

with $\Omega_{V\cdot u} = \Omega_{vv,0} - \frac{\omega_{uv,0}\omega_{vu,0}}{\omega^2_{u,0}}$, implying that Ξ_u, and $\Xi_{V\cdot u}$ are independent conditional on D_ξ.

Proof. By Lemma 3 of Barnichon and Mesters (2020), under the conditions stated in this Lemma, we have

$$\left[K_n^{-1}Z'(y-W\delta_0), K_n^{-1}Z'(W-Z\Pi)\right] = \left[K_n^{-1}Z'u, K_n^{-1}Z'V\right] \Rightarrow \left[\Xi_u, (\Xi_{v1}, \Xi_{v2}, \Xi_{v3})\right] =: \left[\Xi_u, \Xi_V\right], \tag{B.1}$$

where, conditional on D_ξ,

$$\text{vec}(\Xi_u, \Xi_V)|_{D_\xi} \sim N\left(0, \Omega_{uv,0} \otimes \Psi\right). \tag{B.2}$$

When $m < 3$, the elements of Ξ_V must be omitted accordingly.

Now postmultiply equation (B.1) by $R = \begin{bmatrix} 1 & -\frac{\omega_{uv,0}}{\omega^2_{u,0}} \\ 0 & I_m \end{bmatrix}$, and we have

$$\left[K_n^{-1}Z'(y-W\delta_0), K_n^{-1}Z'\left((W-Z\Pi)-(y-W\delta_0)\tfrac{\omega_{uv,0}}{\omega^2_{u,0}}\right)\right] \Rightarrow \left[\Xi_u, \Xi_{V\cdot u}\right],$$

where

$$\text{vec}(\Xi_u, \Xi_{V\cdot u})|_{D_\xi} \sim N\left(0, \begin{bmatrix} \omega^2_{u,0} & 0 \\ 0 & \Omega_{V\cdot u} \end{bmatrix} \otimes \Psi\right)$$

with $\Omega_{V\cdot u} = \Omega_{vv,0} - \frac{\omega_{vu,0}\omega_{uv,0}}{\omega^2_{u,0}}$. This follows because

$$\begin{aligned}
&\text{AVar}\left(\text{vec}\left[K_n^{-1}Z'u, K_n^{-1}Z'\left(V - u\frac{\omega_{uv,0}}{\omega^2_{u,0}}\right)\right]\right) \\
&= \text{Var}\left(\text{vec}\left([\Xi_u, \Xi_V]R\right)\right) \\
&= R'\Omega_{uv,0}R \otimes \Psi \\
&= \begin{bmatrix} \omega^2_{u,0} & 0 \\ 0 & \Omega_{vv,0} - \frac{\omega_{vu,0}\omega_{uv,0}}{\omega^2_{u,0}} \end{bmatrix} \otimes \Psi,
\end{aligned}$$

where the first equality follows from (B.1) and the second equality follows from (B.2). Since Ξ_u and $\Xi_{V\cdot u}$ are uncorrelated and Gaussian, they are independent conditional on D_ξ. □

Lemma B. *Let*

$$\hat{s}_{uw} = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} \hat{u}_t \widehat{w}_s' \kappa((t-s)/b_n) = \frac{1}{n-H} \hat{u}' B_n \widehat{W},$$

and

$$\tilde{s}_{uw} = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} u_t w_s' \kappa((t-s)/b_n) = \frac{1}{n-H} u' B_n W$$

where $\hat{u} = M_Z u$ and $\widehat{W} = M_Z W$.

Given Assumption 1, when $n \to \infty$ with $H/n \to c \in (0,1)$, we have

$$\tilde{s}_{uw} \xrightarrow{p} \omega_{uv,0}$$

and

$$\hat{s}_{uw} = \tilde{s}_{uv} + o_p(1),$$

where $\tilde{s}_{uv} = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} u_t v_s \kappa((t-s)/b_n) = \frac{1}{n-H} u' B_n V$ and B_n is an $(n-H) \times (n-H)$ matrix with (s,t) entry equal to $\kappa((t-s)/b_n)$.

Proof. First, denote $\hat{s}_{uv} = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} \hat{u}_t \hat{v}_s \kappa((t-s)/b_n) = \frac{1}{n-H} \hat{u}' B_n \widehat{V}$, where $\widehat{V} = M_Z V$. Then, since $\widehat{W} = M_Z(Z\Pi + V) = M_Z V = \widehat{V}$, we have $\hat{s}_{uw} = \hat{s}_{uv} = \tilde{s}_{uv} + o_p(1)$, where the last equality holds by Lemma 5 of Barnichon and Mesters (2020).

For $\tilde{s}_{uw}$, we have

$$\begin{aligned}
\tilde{s}_{uw} &= \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} u_t z_s' \Pi \kappa((t-s)/b_n) + \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} u_t v_s' \kappa((t-s)/b_n) \\
&= \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} u_t z_s' \Pi \kappa((t-s)/b_n) + \tilde{s}_{uv}.
\end{aligned}$$

Then, following the same argument in the proof of Lemma 5 of Barnichon and Mesters (2020), the first term converges to its expectations in probability, which is zero because of the orthogonality of the instruments and the error in Assumption 1 (i). Hence, we have $\tilde{s}_{uw} \xrightarrow{p} \tilde{s}_{uv}$, but $\tilde{s}_{uv} \xrightarrow{p} \omega_{uv,0}$ by Lemma 5 of Barnichon and Mesters (2020), concluding that $\tilde{s}_{uw} \xrightarrow{p} \omega_{uv,0}$.

□

Lemma C. *Let* $\tilde{s}_u^2 = \frac{1}{n-H}\sum_{t=H+1}^{n}\sum_{s=H+1}^{n} u_t u_s \kappa((t-s)/b_n) = \frac{1}{n-H}u' B_n u$ *and* $\hat{s}_u^2 = \frac{1}{n-H}\sum_{t=H+1}^{n}\sum_{s=H+1}^{n} \hat{u}_t \hat{u}_s \kappa((t-s)/b_n) = \frac{1}{n-H}\hat{u}' B_n \hat{u}$. *Given Assumption 1, under the null* $H_0 : \delta = \delta_0$ *when* $n \to \infty$ *with* $H/n \to c \in (0,1)$, $K_n^{-1}Z'\left\{(W - Z\Pi) - (y - W\delta_0)\frac{s_{uw}}{s_u^2}\right\}$ *has the same limiting behavior as* $K_n^{-1}Z'\left\{(W - Z\Pi) - (y - W\delta_0)\frac{\omega_{uv,0}}{\omega_{u,0}^2}\right\}$ *so that* $K_n^{-1}Z'\left\{(W - Z\Pi) - (y - W\delta_0)\frac{s_{uw}}{s_u^2}\right\} \Rightarrow \Xi_{V\cdot u}$ *where* $\{s_{uw}, s_u^2\}$ *pair is either* $\{\tilde{s}_{uw}, \tilde{s}_u^2\}$ *or* $\{\hat{s}_{uw}, \hat{s}_u^2\}$

Proof. By adding and subtracting $(y - W\delta_0)\frac{\omega_{uv,0}}{\omega_{u,0}^2}$, we have

$$\begin{aligned} K_n^{-1}Z'\left\{(W - Z\Pi) - (y - W\delta_0)\frac{s_{uw}}{s_u^2}\right\} &= K_n^{-1}Z'\left\{(W - Z\Pi) - (y - W\delta_0)\frac{\omega_{uv,0}}{\omega_{u,0}^2}\right\} \\ &\quad - K_n^{-1}Z'(y - W\delta_0)\left\{(s_{uw} - \omega_{uv,0})\frac{1}{\omega_{u,0}^2} + s_{uw}\left(\frac{1}{s_u^2} - \frac{1}{\omega_{u,0}^2}\right)\right\}. \end{aligned}$$

Then, the second term converges in probability to zero because $K_n^{-1}Z'(y - W\delta_0) = O_p(1)$ by equations (B.1) and (B.2) in Lemma A and $s_{uw} \xrightarrow{p} \omega_{uv,0}$ and $\frac{1}{s_u^2} \xrightarrow{p} \frac{1}{\omega_{u,0}^2}$ by Lemma B and Lemma 5 of Barnichon and Mesters (2020).

Hence, $K_n^{-1}Z'\left\{(W - Z\Pi) - (y - W\delta_0)\frac{s_{uw}}{s_u^2}\right\}$ has the same limiting behavior as $K_n^{-1}Z'\left\{(W - Z\Pi) - (y - W\delta_0)\frac{\omega_{uv,0}}{\omega_{u,0}^2}\right\}$.

Furthermore, by Lemma A, we have $K_n^{-1}Z'\left\{(W - Z\Pi) - (y - W\delta_0)\frac{\omega_{uv,0}}{\omega_{u,0}^2}\right\} \Rightarrow \Xi_{V\cdot u}$, so we have the desired result. □

Lemma D. *If* A *is a* $k \times k$ *nonsingular matrix,* K_n *is a* $k \times k$ *diagonal matrix whose* (j,j) *element is* n^{d_j} *with* d_j *arranged in strictly increasing order (*$d_i < d_j$ *if* $i < j \leq k$*, and* $i, j \in \mathbb{N}$*), and* Π *is a* $k \times m$ *full rank matrix with* $m < k$*. Then, as* $n \to \infty$*, the sequence of a projection matrix* $\{P_{AK_n\Pi}\}$ *converges to a matrix* $\mathcal{P}$ *where* $\mathcal{P}$ *is a projection matrix with* $\text{rank}(\mathcal{P}) = m$.

Proof. Because Π has full column rank m, there exists a nonsingular $m \times m$ matrix Q such that in each column j of $\bar{\Pi} = \Pi Q$, the elements are $\bar{\pi}_{r_j,j} \neq 0$ and $\bar{\pi}_{i,j} = 0$ for all $i > r_j$, and there exist distinct pivot indices $r_1, \ldots, r_m \in \{1, \ldots, k\}$.

Define a normalization matrix $\check{K}_n = \mathrm{diag}(n^{-d_{r_1}}, \ldots, n^{-d_{r_m}})$. Since post-multiplication by a nonsingular matrix does not change the column space, consider

$$B_n \coloneqq AK_n \Pi Q \check{K}_n = A(K_n \bar{\Pi} \check{K}_n) = AG_n$$

where $G_n = K_n \bar{\Pi} \check{K}_n$ and the (i, j) element of G_n is $\bar{\pi}_{i,j} n^{d_i - d_{r_j}}$.

Consider the limit of G_n. We have $\lim_{n\to\infty} G_n = G$, where G has non-zero elements only at (r_j, j), because (a) if $i = r_j$, $(G_n)_{i,j}$ is $\bar{\pi}_{r_j,j} n^0 \to \bar{\pi}_{r_j,j} \neq 0$; (b) if $i > r_j$, $\bar{\pi}_{i,j} = 0$, so $(G_n)_{i,j} = 0$; and (c) if $i < r_j$, since $d_i < d_{r_j}$, $n^{d_i - d_{r_j}} \to 0$ and we have $(G_n)_{i,j} \to 0$, as $n \to \infty$. This implies that $B_\infty \coloneqq \lim_{n\to\infty} B_n = AG$. Since each column of G contains exactly one nonzero element and the pivot indices are distinct, G has full column rank m.

Then, since A is nonsingular and G has full column rank m, B_∞ also has full column rank m. Because the projection matrix is a continuous transformation for a sequence of full-rank matrices with constant rank, we have

$$\lim_{n\to\infty} P_{AK_n\Pi} = \lim_{n\to\infty} P_{AK_n\Pi Q\check{K}_n} = \lim_{n\to\infty} P_{AK_n\bar{\Pi}\check{K}_n} = \lim_{n\to\infty} P_{AG_n} = P_{AG} = P_{B_\infty} \eqqcolon \mathcal{P}.$$

Since B_∞ has rank m, the projection matrix $\mathcal{P}$ also has rank m.

□

Lemma 4 of Barnichon and Mesters (2020). *Given Assumption 1 we have when $n \to \infty$ with $H/n \to c \in (0, 1)$ that*

$$K_n^{-1} \sum_{t=H+1}^{n} z_t z_t' K_n^{-1} \Rightarrow \int_c^1 D_\xi(a) D_\xi'(a) da.$$

Some remarks on Lemma 4: See Lemma A for the defininion of K_n and $D_\xi(a)$. In this paper, I denote $\Psi \coloneqq \int_c^1 D_\xi(a) D_\xi'(a) da$.

Lemma 5 of Barnichon and Mesters (2020). *Given Assumption 1, let*

$$\widehat{S}_{uv} = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} (\hat{u}_t, \hat{v}_t')'(\hat{u}_s, \hat{v}_s')\kappa((t-s)/b_n) = \frac{1}{n-H}(\hat{u} : \widehat{V})'B_n(\hat{u} : \widehat{V})$$

Also, define

$$\widetilde{S}_{uv} = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} (u_t, v_t')'(u_s, v_s')\kappa((t-s)/b_n) = \frac{1}{n-H}(u : V)'B_n(u : V).$$

Under the conditions of Assumption 1 we have when $n \to \infty$ with $H/n \to c \in (0,1)$ that

$$\widetilde{S}_{uv} \xrightarrow{p} \Omega_{uv,0}$$

and

$$\widehat{S}_{uv} = \widetilde{S}_{uv} + o_p(1).$$

B.2 Main proof of Theorem 1

With these lemmas, we are ready to prove Theorem 1.

Proof of Theorem 1. I first prove the results for K_a^r. For notational convenience, I suppress the dependence of $\widetilde{\Pi}(\delta_0)$ and $\widetilde{W}(\delta_0)$ and so forth on δ_0 in the following proof, unless otherwise confusing.

First, rewrite the KLM statistic $K_a^r(\delta_0)$ as:

$$\begin{aligned} K_a^r(\delta_0) &= \frac{1}{\tilde{s}_u^2}(y - W\delta_0)'P_{\widetilde{W}}(y - W\delta_0) \\ &= \frac{1}{\tilde{s}_u^2} u'ZK_n^{-1}K_n\widetilde{\Pi}\left[\widetilde{\Pi}'K_n(K_n^{-1}Z'ZK_n^{-1})K_n\widetilde{\Pi}\right]^{-1}\widetilde{\Pi}'K_nK_n^{-1}Z'u \qquad \text{(B.3)} \\ &= \frac{1}{\tilde{s}_u^2}(K_n^{-1}Z'u)'D_n\left[D_n'(K_n^{-1}Z'ZK_n^{-1})D_n\right]^{-1}D_n'(K_n^{-1}Z'u) \end{aligned}$$

where

$$\widetilde{W} = Z\widetilde{\Pi}, \quad \widetilde{\Pi} = (Z'Z)^{-1}Z'\left[W - (y - W\delta_0)\frac{\tilde{s}_{uw}}{\tilde{s}_u^2}\right]$$

and

$$\begin{aligned} D_n &:= K_n\widetilde{\Pi} \\ &= K_n\Pi + (K_n^{-1}Z'ZK_n^{-1})^{-1}K_n^{-1}Z'\left[(W - Z\Pi) - (y - W\delta_0)\frac{\tilde{s}_{uw}}{\tilde{s}_u^2}\right] \end{aligned} \tag{B.4}$$

because

$$\widetilde{\Pi} = \Pi + (Z'Z)^{-1}Z'\left[(W - Z\Pi) - (y - W\delta_0)\frac{\tilde{s}_{uw}}{\tilde{s}_u^2}\right].$$

I consider three cases for the instrument strength: relevant, weakly relevant, and irrelevant.

(i) The instruments are relevant such that Π is fixed and has full rank.

When the instruments are relevant, Π is fixed and has full rank, so we have

$$D_n = K_n\Pi + O_p(1)$$

for equation (B.4), because $(K_n^{-1}Z'ZK_n^{-1})^{-1}K_n^{-1}Z'\left[(W - Z\Pi) - (y - W\delta_0)\frac{\tilde{s}_{uw}}{\tilde{s}_u^2}\right] \Rightarrow \Psi^{-1}\Xi_{V\cdot u}$ by Lemma 4 of Barnichon and Mesters (2020) and Lemma A. Hence, for sufficiently large n, we have that $D_n \approx K_n\Pi =: \mathcal{D}_n$, because $K_n = \text{diag}(O(n), O(n^2), O(n^3))$ and thus $K_n\Pi$ has the dominant effect on D_n.

Then, by Lemmas 4 and 5 of Barnichon and Mesters (2020) and Lemmas A, C, D, the KLM statistic (B.3) converges to

$$K_a^r(\delta_0) \Rightarrow \left(\frac{\Psi^{-1/2}\Xi_u}{\omega_{u,0}}\right)' \mathcal{P}\left(\frac{\Psi^{-1/2}\Xi_u}{\omega_{u,0}}\right)$$

where $\mathcal{P} := \lim_{n\to\infty} P_{\Psi^{1/2}\mathcal{D}_n}$ exists and its rank is m by Lemma D because $\Psi^{1/2}$ is nonsingular almost surely. Since $\left(\frac{\Psi^{-1/2}\Xi_u}{\omega_{u,0}}\right)$ follows a standard multivariate normal distribution conditional on Ψ, we have $K_a^r(\delta_0)|_\Psi \Rightarrow \chi^2(m)$. Since the conditional limiting distribution does not depend on Ψ, we have $K_a^r(\delta_0) \Rightarrow \chi^2(m)$ unconditionally.

(ii) The instruments are weak such that $\Pi = \Pi_n = K_n^{-1}C$ with C being fixed and full rank.

Assume $\Pi = K_n^{-1}C$. Then, by Lemma 4 of Barnichon and Mesters (2020) and Lemma A, we have

$$D_n \Rightarrow C + \Psi^{-1}\Xi_{V\cdot u}$$

for equation (B.4). Then, by Lemmas 4 and 5 of Barnichon and Mesters (2020) and Lemmas A and C, the KLM statistic (B.3) converges to

$$\begin{aligned} K_a^r(\delta_0) &\Rightarrow \frac{1}{\omega_{u,0}^2}\Xi_u'(C + \Psi^{-1}\Xi_{V\cdot u})\left((C + \Psi^{-1}\Xi_{V\cdot u})'\Psi^{-1}(C + \Psi^{-1}\Xi_{V\cdot u})\right)^{-1}(C + \Psi^{-1}\Xi_{V\cdot u})'\Xi_u \\ &= \left(\frac{\Psi^{-1/2}\Xi_u}{\omega_{u,0}}\right)' P_{\Psi^{1/2}(C+\Psi^{-1}\Xi_{V\cdot u})}\left(\frac{\Psi^{-1/2}\Xi_u}{\omega_{u,0}}\right). \end{aligned}$$

This implies $K_a^r(\delta_0)|_\Psi \Rightarrow \chi^2(m)$ because Ξ_u and $\Xi_{V\cdot u}$ are independent by Lemma A; $\text{rank}(\Psi^{1/2}(C + \Psi^{-1}\Xi_{V\cdot u})) = m$ since C has full rank, $\Xi_{V\cdot u}$ has full rank almost surely because it is Gaussian, and Ψ is non-singular almost surely. Therefore, also unconditionally, we have $K_a^r(\delta_0) \Rightarrow \chi^2(m)$.

(iii) The instruments are irrelevant such that $\Pi = 0$.

Assume $\Pi = 0$. Then by Lemma 4 of Barnichon and Mesters (2020) and Lemma A, we have

$$D_n \Rightarrow \Psi^{-1}\Xi_{V\cdot u}.$$

for equation (B.4). Then, by Lemmas 4 and 5 of Barnichon and Mesters (2020) and Lemmas A and C, the KLM statistic (B.3) converges to

$$\begin{aligned} K_a^r(\delta_0) &\Rightarrow \frac{1}{\omega_{u,0}^2}\Xi_u'(\Psi^{-1}\Xi_{V\cdot u})\left((\Psi^{-1}\Xi_{V\cdot u})'\Psi^{-1}(\Psi^{-1}\Xi_{V\cdot u})\right)^{-1}(\Psi^{-1}\Xi_{V\cdot u})'\Xi_u \\ &= \left(\frac{\Psi^{-1/2}\Xi_u}{\omega_{u,0}}\right)' P_{\Psi^{-1/2}\Xi_{V\cdot u}}\left(\frac{\Psi^{-1/2}\Xi_u}{\omega_{u,0}}\right). \end{aligned}$$

Since Ξ_u and $\Xi_{V\cdot u}$ are independent by Lemma A; $\text{rank}(\Psi^{-1/2}\Xi_{V\cdot u}) = m$ because Ψ is nonsingular almost surely and $\Xi_{V\cdot u}$ has full rank almost surely; we have that $K_a^r(\delta_0)|_\Psi \Rightarrow \chi^2(m)$, implying that unconditionally $K_a^r(\delta_0) \Rightarrow \chi^2(m)$.

Proof for K_a

The only difference between K_a and K_a^r is the use of the restricted long-run variance

$(\hat{s}_{uw}, \hat{s}_u^2)$, but their consistency by Lemma B ensures Lemma C holds even with the unrestricted long-run variance, and thus the result on K_a follows by the same argument as K_a^r. □

C Additional simulation results

Tables C.1 and C.2 summarize the simulation results discussed in Section 3 at significance levels of $\alpha = 0.1$ and $\alpha = 0.32$, respectively. These results confirm the robustness of the findings reported in the main paper. First, AR_a^r and K_a^r with χ^2 critical values maintain an accurate finite-sample size. Second, AR_a and K_a with χ^2 critical values over-reject the null hypothesis. Third, employing the fixed-b critical values of Sun (2014) improves the finite-sample performance of AR_a and K_a. Fourth, both the AR and KLM statistics without Almon parameterization over-reject the null.

			With $\gamma_b + \gamma_f = 1$												Without $\gamma_b + \gamma_f = 1$			
			χ^2						fixed-b						χ^2		fixed-b	
T	σ_i	ρ	AR_a^r	AR_a	K_a^r	K_a	K^r	K	AR_a^r	AR_a	K_a^r	K_a	K^r	K	AR_a^r	AR_a	AR_a^r	AR_a
100	0.10	0.0	10.9	16.9	8.2	14.9	93	95	3.6	8.5	3.9	9.2	87	92	10.9	16.3	3.6	8.1
100	0.25	0.0	11.0	17.0	8.5	14.8	87	88	3.5	8.5	4.3	9.2	79	82	11.3	16.6	4.0	8.3
100	0.50	0.0	11.6	16.4	9.6	14.8	75	71	3.8	8.0	5.1	9.3	64	63	11.2	16.2	3.7	8.2
100	1.00	0.0	12.5	15.5	10.9	13.8	58	44	5.1	7.6	6.2	8.8	48	35	11.0	15.9	3.6	7.5
100	0.10	0.5	11.8	19.4	7.9	16.1	17	35	3.9	10.4	3.8	10.6	9	23	11.8	19.6	3.9	10.5
100	0.25	0.5	11.1	18.4	7.6	15.4	17	33	3.6	10.1	3.6	10.1	9	21	11.8	19.0	3.6	10.3
100	0.50	0.5	11.7	18.4	8.0	15.6	19	31	3.8	10.1	3.9	10.5	11	21	11.7	19.1	3.6	10.1
100	1.00	0.5	12.0	17.8	9.2	15.9	23	31	4.6	9.6	4.9	10.2	14	22	11.4	18.3	3.7	9.7
200	0.10	0.0	10.5	13.5	8.1	11.3	92	93	6.7	9.3	5.9	8.9	89	91	10.7	13.4	6.6	9.1
200	0.25	0.0	10.5	13.6	8.6	11.6	81	82	6.5	9.2	6.3	9.2	77	78	10.7	13.4	6.6	9.1
200	0.50	0.0	10.8	13.3	9.0	11.6	61	57	6.5	9.0	6.8	9.1	55	52	10.6	13.1	6.8	9.0
200	1.00	0.0	11.1	12.4	10.0	11.6	41	29	7.2	8.7	7.7	8.9	36	25	10.6	12.8	6.7	9.0
200	0.10	0.5	10.9	14.8	7.7	11.3	15	23	6.7	10.3	5.5	9.1	10	17	10.6	14.3	6.9	10.0
200	0.25	0.5	11.0	14.4	8.1	11.8	16	23	7.0	10.3	6.1	9.4	11	17	10.8	14.3	6.8	10.2
200	0.50	0.5	11.4	14.7	7.9	11.4	16	21	7.4	10.7	5.8	9.1	12	17	11.3	14.8	7.1	10.5
200	1.00	0.5	11.4	14.3	9.4	12.1	18	20	7.2	10.2	7.0	9.8	14	16	11.2	14.3	7.2	10.4

Note: This table reports the rejection frequencies (%) for $H_0 : \delta = \delta_0$ at a significance level of $\alpha = 0.10$, based on 10,000 simulations. The left panel reports results with the common restriction $\gamma_g + \gamma_f = 1$, while the right panel reports results without the restriction. Each row corresponds to a specific sample size $T \in \{100, 200\}$, standard deviation of the shock $\sigma_i \in \{0.1, 0.25, 0.5, 1.0\}$, and autocorrelation of the disturbance $\rho \in \{0.0, 0.5\}$. Tests are based on the Almon-restricted AR and KLM and the standard KLM statistics with restricted long-run variance (AR_a^r, K_a^r, K^r) or versions with unrestricted long-run variance (AR_a, K_a, K), using critical values either from the χ^2 distribution or from the fixed-b asymptotics of Sun (2014). Under the restriction, the degrees of freedom for the AR and KLM statistics are three and two, respectively. Without the restriction, the degrees of freedom are three.

Table C.1: Rejection frequencies (%) for $\alpha = 0.10$

			With $\gamma_b + \gamma_f = 1$												Without $\gamma_b + \gamma_f = 1$			
			χ^2						fixed-b						χ^2		fixed-b	
T	σ_i	ρ	AR_a^r	AR_a	K_a^r	K_a	K^r	K	AR_a^r	AR_a	K_a^r	K_a	K^r	K	AR_a^r	AR_a	AR_a^r	AR_a
100	0.10	0.0	36.0	39.6	30.0	37.2	99	99	25.1	29.9	24.1	31.4	99	99	35.8	39.4	24.8	29.2
100	0.25	0.0	35.7	39.4	30.9	37.5	97	97	25.1	29.7	24.2	31.6	96	96	36.5	39.8	25.6	30.0
100	0.50	0.0	35.7	39.6	31.0	36.8	92	87	25.1	29.6	24.8	31.1	89	84	36.2	39.5	25.1	29.7
100	1.00	0.0	35.5	37.9	32.7	36.5	81	67	25.9	27.8	26.7	30.3	77	62	35.7	38.8	25.3	29.3
100	0.10	0.5	38.4	42.7	29.2	37.9	61	74	27.2	33.4	23.3	32.1	50	67	38.2	42.4	27.4	33.0
100	0.25	0.5	37.3	41.6	29.3	37.7	57	69	25.7	31.9	22.9	31.6	47	61	37.7	42.1	26.5	32.4
100	0.50	0.5	37.6	41.8	30.0	37.5	52	62	26.2	31.8	24.1	32.1	45	55	37.6	42.0	26.6	32.7
100	1.00	0.5	37.5	40.9	31.4	38.4	52	59	26.1	30.9	25.5	32.2	45	53	37.0	40.9	26.2	31.8
200	0.10	0.0	34.4	36.5	29.4	33.6	99	99	28.7	31.4	26.4	30.2	98	98	34.4	36.1	29.2	31.4
200	0.25	0.0	33.8	35.8	29.5	33.0	95	94	28.7	30.9	26.7	30.2	94	93	34.1	36.2	29.1	31.3
200	0.50	0.0	33.8	35.7	30.8	33.7	83	78	28.4	30.7	27.5	30.9	81	76	33.6	35.5	28.3	30.4
200	1.00	0.0	33.5	34.4	31.3	33.2	68	55	28.3	29.3	28.2	30.2	66	52	34.0	35.6	28.6	30.7
200	0.10	0.5	35.6	38.1	27.8	32.3	60	67	30.1	33.0	24.8	29.5	55	63	34.6	37.2	29.4	32.1
200	0.25	0.5	34.6	36.9	28.2	32.4	53	59	29.4	32.0	25.5	29.4	49	55	34.6	36.9	29.4	31.9
200	0.50	0.5	35.2	37.5	28.8	32.8	46	51	29.9	32.4	25.6	29.8	42	48	35.4	37.7	29.8	32.5
200	1.00	0.5	35.6	37.7	31.0	34.0	43	47	30.3	32.6	28.0	30.9	40	44	34.9	36.7	29.4	31.9

Note: This table reports the rejection frequencies (%) for $H_0 : \delta = \delta_0$ at a significance level of $\alpha = 0.32$, based on 10,000 simulations. The left panel reports results with the common restriction $\gamma_g + \gamma_f = 1$, while the right panel reports results without the restriction. Each row corresponds to a specific sample size $T \in \{100, 200\}$, standard deviation of the shock $\sigma_i \in \{0.1, 0.25, 0.5, 1.0\}$, and autocorrelation of the disturbance $\rho \in \{0.0, 0.5\}$. Tests are based on the Almon-restricted AR and KLM and the standard KLM statistics with restricted long-run variance (AR_a^r, K_a^r, K^r) or versions with unrestricted long-run variance (AR_a, K_a, K), using critical values either from the χ^2 distribution or from the fixed-b asymptotics of Sun (2014). Under the restriction, the degrees of freedom for the AR and KLM statistics are three and two, respectively. Without the restriction, the degrees of freedom are three.

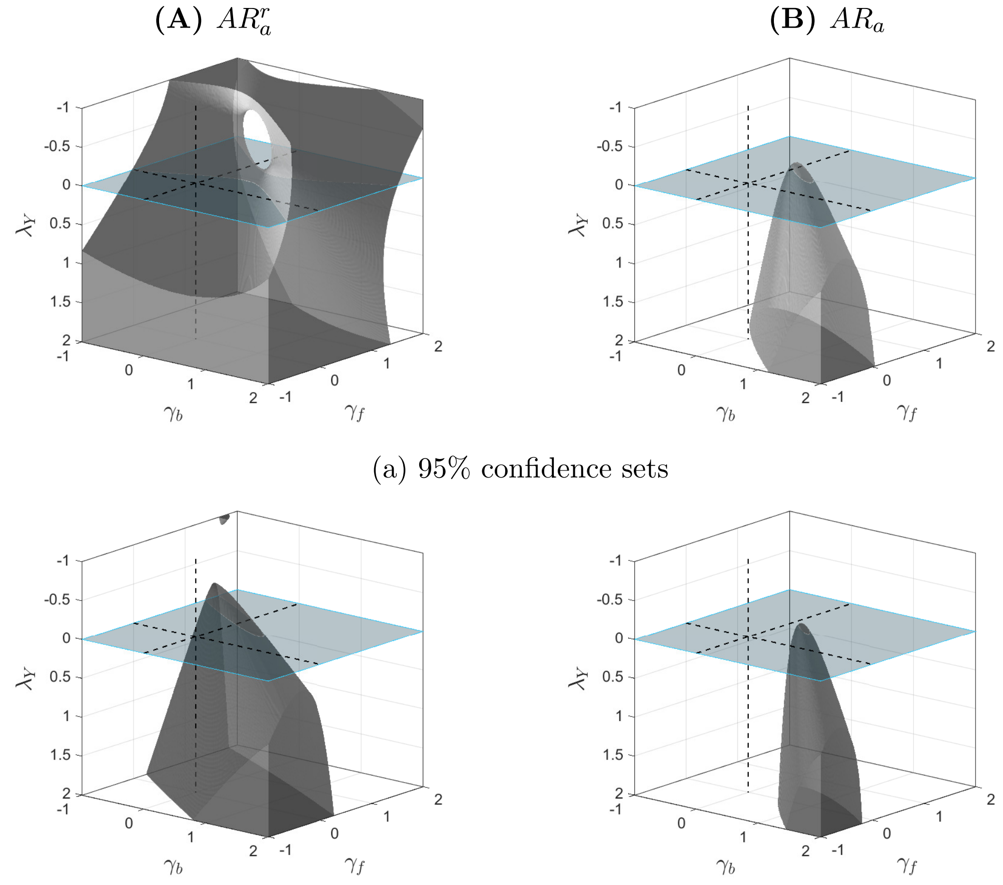
Table C.2: Rejection frequencies (%) for $\alpha = 0.32$

D Additional results on the Phillips curve

D.1 Phillips curve in 1969–2007 with the output gap

Figure D.1 presents the robust confidence sets for the Phillips curve over the 1969—2007 period, identified by the Romer-Romer proxy with the *output* gap as the forcing variable. Panels (a) and (b) plot the 95 and 90%, confidence sets over three parameters $\delta = [\lambda_Y, \gamma_b, \gamma_f]$, respectively. The gray areas indicate the parameter values not rejected by either AR_a^r (left) or AR_a (right) using $\chi^2(3)$ critical values. The plane corresponds to a completely flat Phillips curve: $\lambda_Y = 0$. Figure D.2 shows the 95, 90, and 68% confidence sets for the two parameters γ_f and λ_Y under the common restriction $\gamma_b + \gamma_f = 1$.

The figure confirms the findings presented in the main text, which used the unemployment gap as the forcing variable. First, the confidence sets for $\delta = [\lambda_Y, \gamma_b, \gamma_f]$ are substantially wider with AR_a^r than with AR_a. Notably, AR_a^r does not reject a completely flat Phillips curve over a wider range of γ_b and γ_f at the 5% significance level. Second, a similar observation holds under the common restriction. Third, the confidence sets in Panel (b) of Figure D.2 are wider than the results in Figure II (B) of Barnichon and Mesters (2020), which is replicated in Panel (c) of Figure D.2, even though I use the same specification and datasets. This discrepancy arises because they used $\chi^2(2)$ critical values rather than the correct $\chi^2(3)$ critical values. These results are robust when AR_a with fixed-b critical values is used, as shown in Figure F.1.

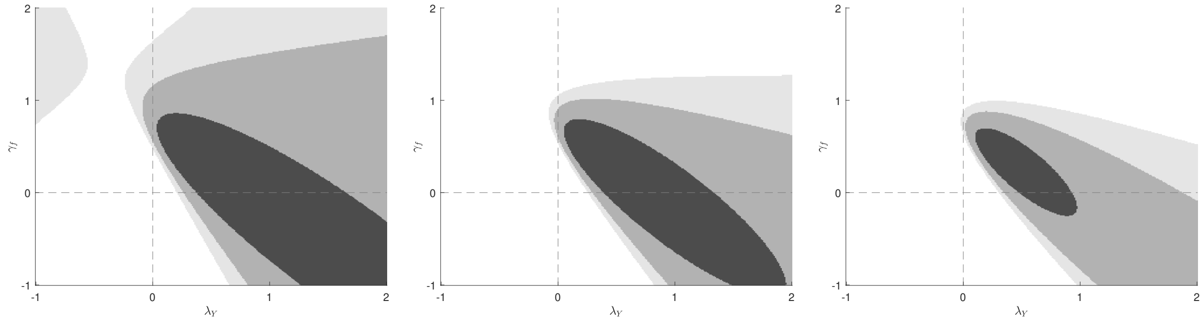


(a) 95% confidence sets

(b) 90% confidence sets

Figure D.1: Phillips curve with output gap identified by the Romer-Romer proxy

Note: Panels (a) and (b) plot the 95 and 90% confidence sets over three parameters $\delta = [\lambda_Y, \gamma_b, \gamma_f]$, respectively, where the gray areas are the sets that are not rejected either by AR_a^r (on the left) or AR_a (on the right) using $\chi^2(3)$ critical values. The plane represents the area of a completely flat Phillips curve: $\lambda_Y = 0$.



(a) AR_a^r with $\chi^2(3)$ (b) AR_a with $\chi^2(3)$ (c) Barnichon and Mesters

Figure D.2: Phillips curve with output gap identified by the Romer-Romer proxy under $\gamma_b + \gamma_f = 1$.

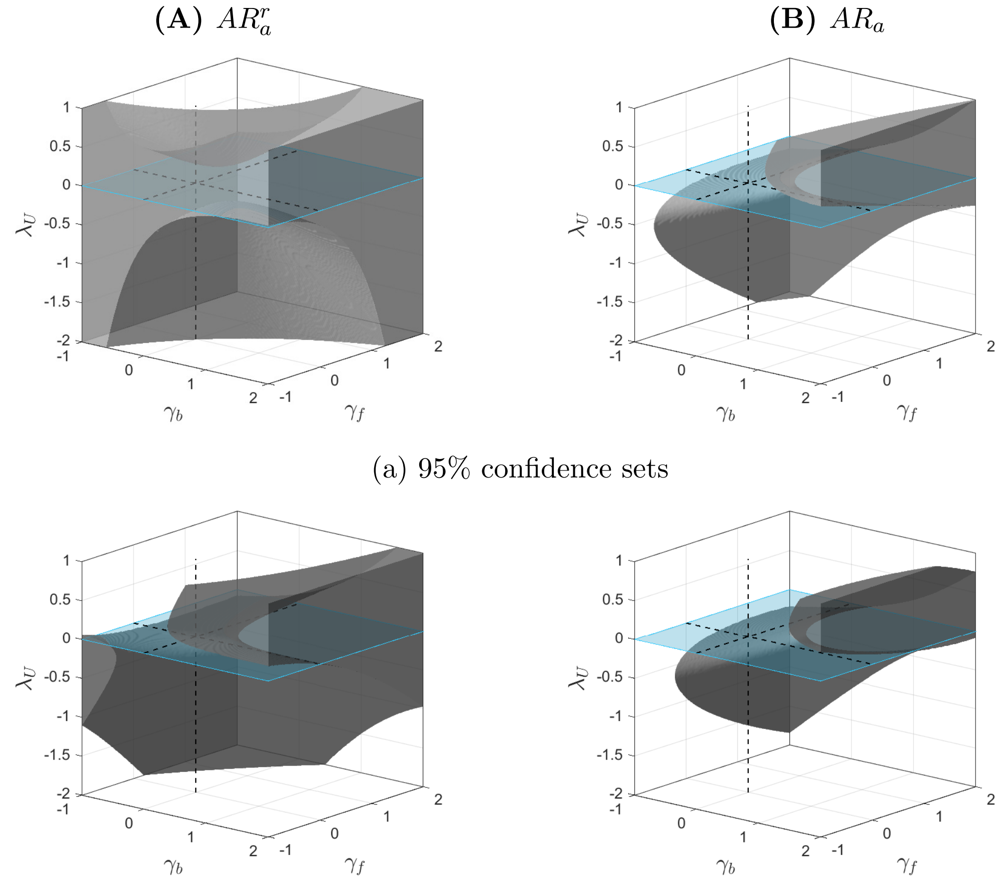
Note: This figure shows the 95% (light gray), 90% (gray), and 68% (black) confidence sets for the two parameters γ_f and λ_U, under the common restriction $\gamma_b + \gamma_f = 1$. Panels (a) and (b) use AR_a^r and AR_a, respectively, with the $\chi^2(3)$ critical values. Panel (c) replicates Figure II (B) of Barnichon and Mesters (2020), which uses AR_a with $\chi^2(2)$ critical values

D.2 Phillips curve in 1990–2017 identified by the HFI proxy

Figure D.3 presents empirical results using high-frequency-identified monetary surprises as instruments for the 1990–2017 sample period, with the unemployment gap as the forcing variable. Following Barnichon and Mesters (2020), I use $H = 24$ lags of shock proxies to construct the Almon-parameterized instruments. Panels (a) and (b) plot the 95 and 90% confidence sets over the three parameters $\delta = [\lambda_U, \gamma_b, \gamma_f]$, respectively. The gray areas represent the parameter values not rejected by either AR_a^r (left) or AR_a (right) using $\chi^2(3)$ critical values. The plane corresponds to the area of a completely flat Phillips curve ($\lambda_U = 0$). Figure D.4 shows the 95, 90, and 68% confidence sets for the two parameters γ_f and λ_U under the common restriction $\gamma_b + \gamma_f = 1$.

We observe results similar to those obtained with the Romer-Romer proxy. First, the confidence sets are wider for AR_a^r, and they include broader regions where parameter values are inconsistent with theoretical predictions, such as a positive λ_U or values of γ_f and γ_b that are negative or exceed unity. Second, the results are similar when the restriction $\gamma_b + \gamma_f = 1$ is imposed. Third, the confidence sets with AR_a under the common restriction (Panel (b) of Figure D.4) are wider than those reported by Barnichon and Mesters (2020) in their Figure V, which is replicated in Panel (c) of Figure D.4, for the reason discussed previously: the use of $\chi^2(2)$, not $\chi^2(3)$, critical values.

Figures D.5 and D.6 confirm that the results are similar when the forcing variable is the output gap. These results are robust when AR_a with fixed-b critical values is used, as shown in Figure F.3.

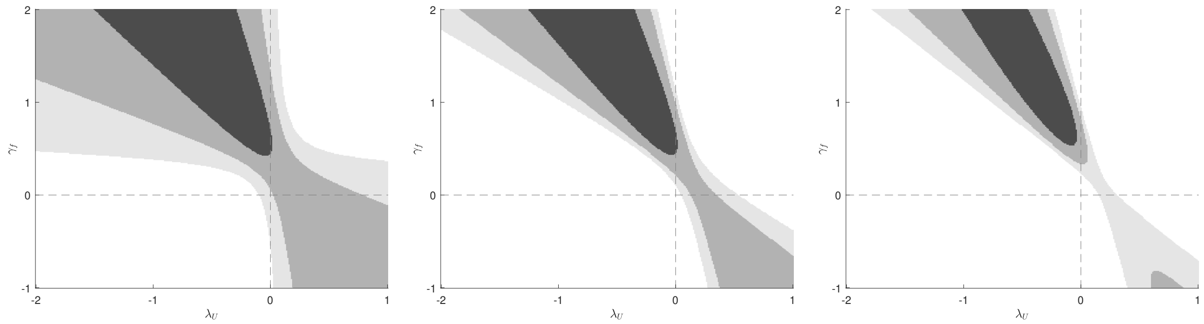


(a) 95% confidence sets

(b) 90% confidence sets

Figure D.3: Phillips curve with unemployment gap identified by the high-frequency identified proxy

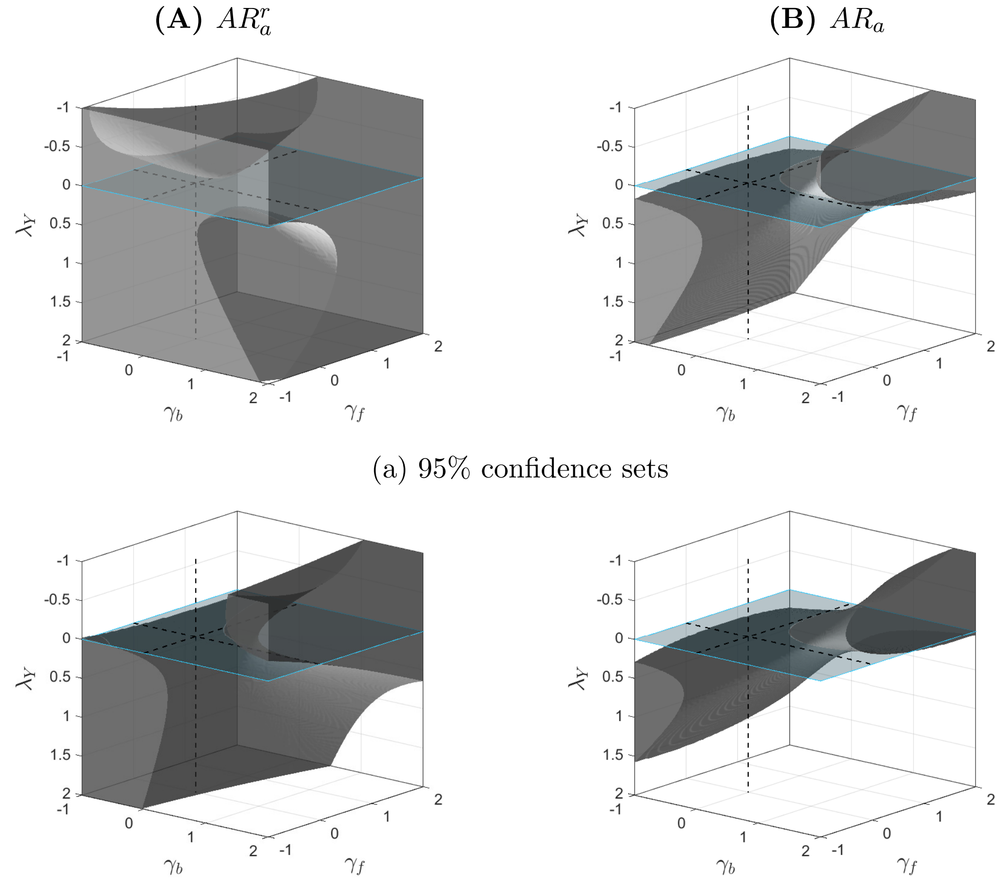
Note: Panels (a) and (b) plot the 95 and 90% confidence sets over three parameters $\delta = [\lambda_U, \gamma_b, \gamma_f]$, respectively, where the gray areas are the sets that are not rejected either by AR_a^r (on the left) or AR_a (on the right) using $\chi^2(3)$ critical values. The plane represents the area of a completely flat Phillips curve: $\lambda_U = 0$.



(a) AR_a^r with $\chi^2(3)$ (b) AR_a with $\chi^2(3)$ (c) Barnichon and Mesters

Figure D.4: Phillips curve with unemployment gap identified by the HFI proxy under $\gamma_b + \gamma_f = 1$.

Note: This figure shows the 95% (light gray), 90% (gray), and 68% (black) confidence sets for the two parameters γ_f and λ_U, under the common restriction $\gamma_b + \gamma_f = 1$. Panels (a) and (b) use AR_a^r and AR_a, respectively, with the $\chi^2(3)$ critical values. Panel (c) replicates Figure V (A) of Barnichon and Mesters (2020), which uses AR_a with $\chi^2(2)$ critical values

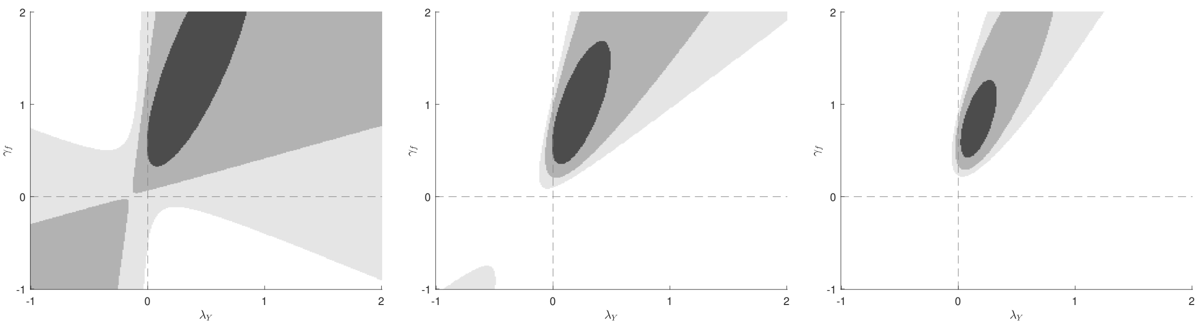


(a) 95% confidence sets

(b) 90% confidence sets

Figure D.5: Phillips curve with output gap identified by the high-frequency identified proxy

Note: Panels (a) and (b) plot the 95 and 90% confidence sets over three parameters $\delta = [\lambda_Y, \gamma_b, \gamma_f]$, respectively, where the gray areas are the sets that are not rejected either by AR_a^r (on the left) or AR_a (on the right) using $\chi^2(3)$ critical values. The plane represents the area of a completely flat Phillips curve: $\lambda_Y = 0$.



(a) AR_a^r with $\chi^2(3)$ (b) AR_a with $\chi^2(3)$ (c) Barnichon and Mesters

Figure D.6: Phillips curve with output gap identified by the HFI proxy under $\gamma_b + \gamma_f = 1$.

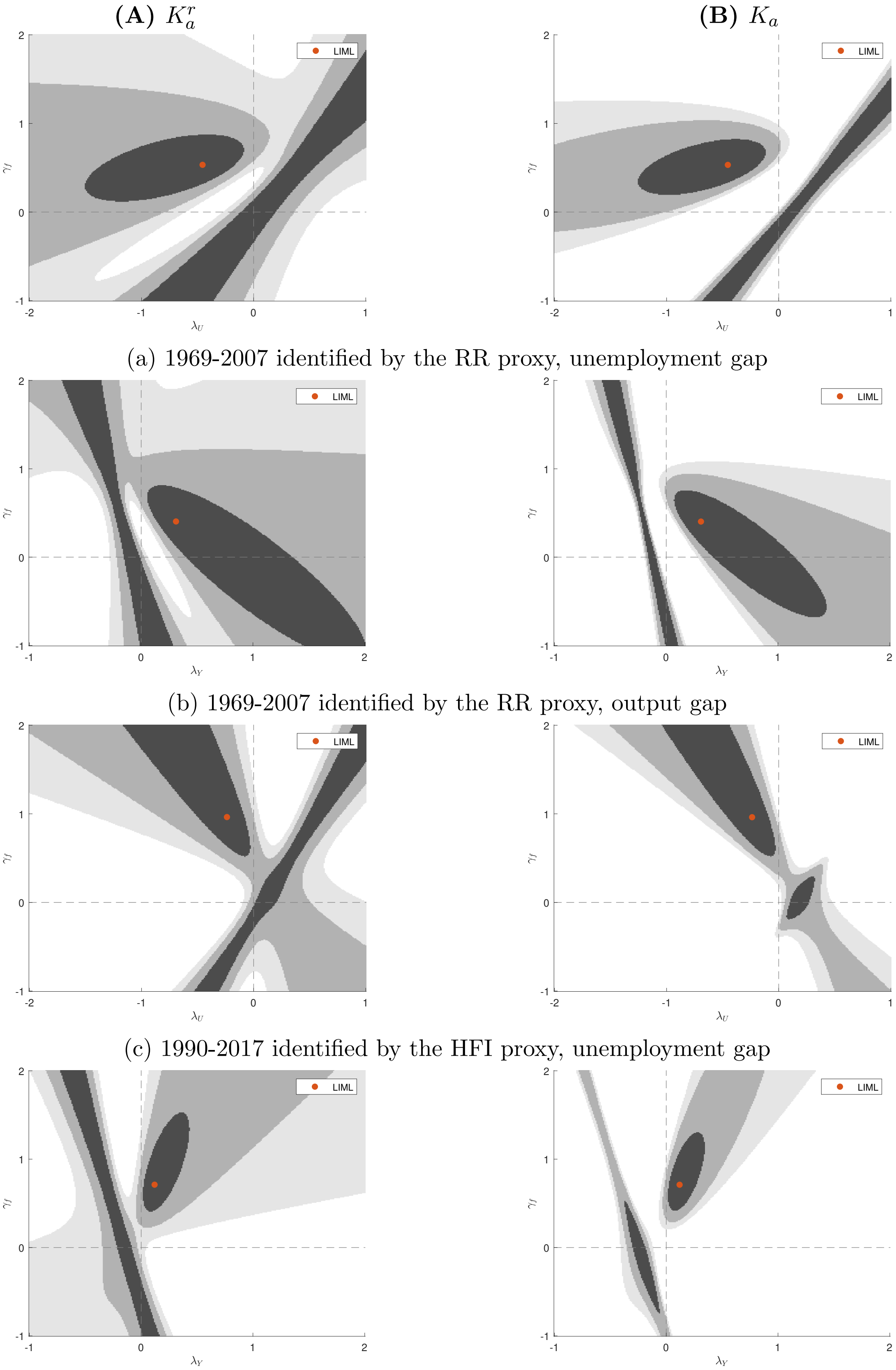
Note: This figure shows the 95% (light gray), 90% (gray), and 68% (black) confidence sets for the two parameters γ_f and λ_U, under the common restriction $\gamma_b + \gamma_f = 1$. Panels (a) and (b) use AR_a^r and AR_a, respectively, with the $\chi^2(3)$ critical values. Panel (c) replicates Figure V (B) of Barnichon and Mesters (2020), which uses AR_a with $\chi^2(2)$ critical values

D.3 Phillips curve for both periods with K_a^r and K_a

In the main paper, I presented the results of the Phillips curve under the common restriction $\gamma_b + \gamma_f = 1$ using the AR statistic to allow direct comparison with Barnichon and Mesters (2020). Still, the model is overidentified, so it is generally recommended to use the Almon-parameterized KLM statistic.

Figure D.7 shows the confidence sets of the Phillips curve under the common restriction $\gamma_b + \gamma_f = 1$ with the KLM statistics. As mentioned in the main paper, since the KLM statistic is a quadratic form of the score of the concentrated log-likelihood, it achieves zero not only at the maximum but also at the minimum. Therefore, we should ignore the finite subsets of the confidence sets that do not contain the LIML estimate.

We observe results similar to those in the main paper. Particularly, the confidence sets obtained by K_a^r are wider than those by K_a, and the quantitative conclusion may change in some cases. For instance, Panel (a) shows that we may conclude that the slope of the Phillips curve is statistically significant at the 10% significance level with K_a, while this is not true with the size-accurate statistic K_a^r.



(a) 1969-2007 identified by the RR proxy, unemployment gap

(b) 1969-2007 identified by the RR proxy, output gap

(c) 1990-2017 identified by the HFI proxy, unemployment gap

(d) 1990-2017 identified by the HFI proxy, output gap

Figure D.7: With $\gamma_b + \gamma_f = 1$. Confidence sets by KLM statistic and $\chi^2(2)$ critical values

Note: This figure shows the 95% (light gray), 90% (gray), and 68% (black) confidence sets for the two parameters γ_f and λ_U or λ_Y, under the common restriction $\gamma_b+\gamma_f = 1$, using K_a^r (left) or K_a (right) with $\chi^2(2)$ critical values. The dot represents the limited information maximum likelihood (LIML) estimate.

E Additional results on the fiscal policy rules

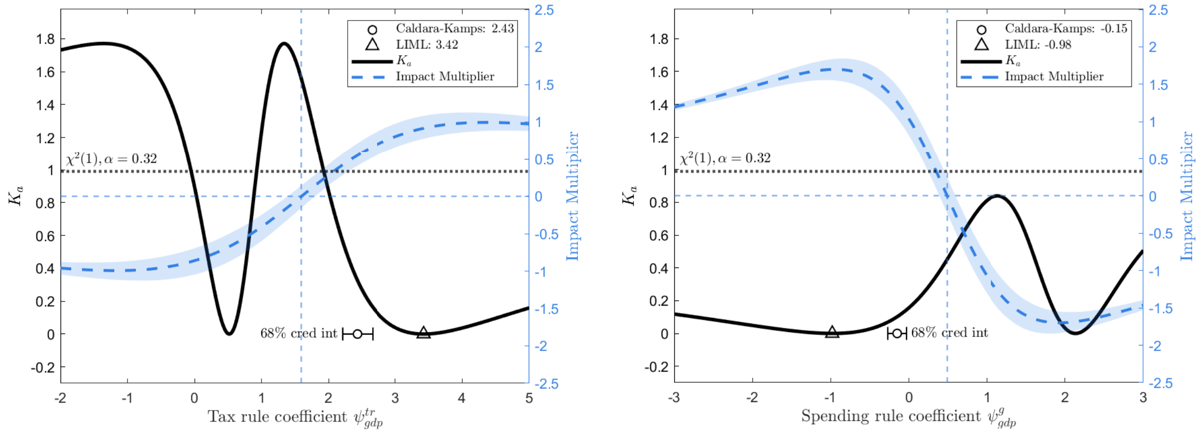
This section presents additional results for the simple and general fiscal rules using K_a with $\chi^2(1)$ critical values, which were shown to over-reject the null in finite samples via the Monte Carlo simulations. The purpose of this section is to complement the argument about the Phillips curve that using size-distorted statistics may affect empirical conclusions. Regarding the fiscal rules, the differences between $\{K_a, K_a^r\}$ and $\{AR_a^r, AR_a\}$ are small, and the quantitative conclusions are similar. Still, we observe that using K_a or AR_a with the standard χ^2 critical values tends to produce tighter confidence sets than K_a^r, implying that using the size-accurate statistics is important to avoid overly narrow confidence sets.

E.1 Simple fiscal rules and multipliers

Figure E.1 presents the K_a statistic (solid line) and the impact multiplier (dashed line) calculated by equation (13) as a function of the systematic response of fiscal policy to output for the simple tax (ψ_{gdp}^{tr}) and spending (ψ_{gdp}^{g}) rules in equation (12). The dashed vertical line indicates the values of the systematic response for which the impact multiplier is zero. The shaded area around the impact multiplier is the pointwise confidence interval obtained from $10,000$ bootstrap samples, while the dotted horizontal line represents the critical value for K_a based on the $\chi^2(1)$ distribution. Following Caldara and Kamps (2017), I adopt a 32% significance level.

In the case of the tax rule, the confidence set becomes slightly narrower with K_a than with K_a^r, but this difference is not large enough to change the quantitative conclusion of the main paper. Nonetheless, similar to the Phillips curve in Section 4, this result implies that using size-distorted statistics may induce overly narrow confidence sets in practice.

In the case of the spending rule, the empirical conclusion is not affected whether K_a^r or K_a is used, as no values of ψ_{gdp}^{g} are rejected for $\psi_{gdp}^{g} \in [-3, 3]$.



(a) Simple tax rule

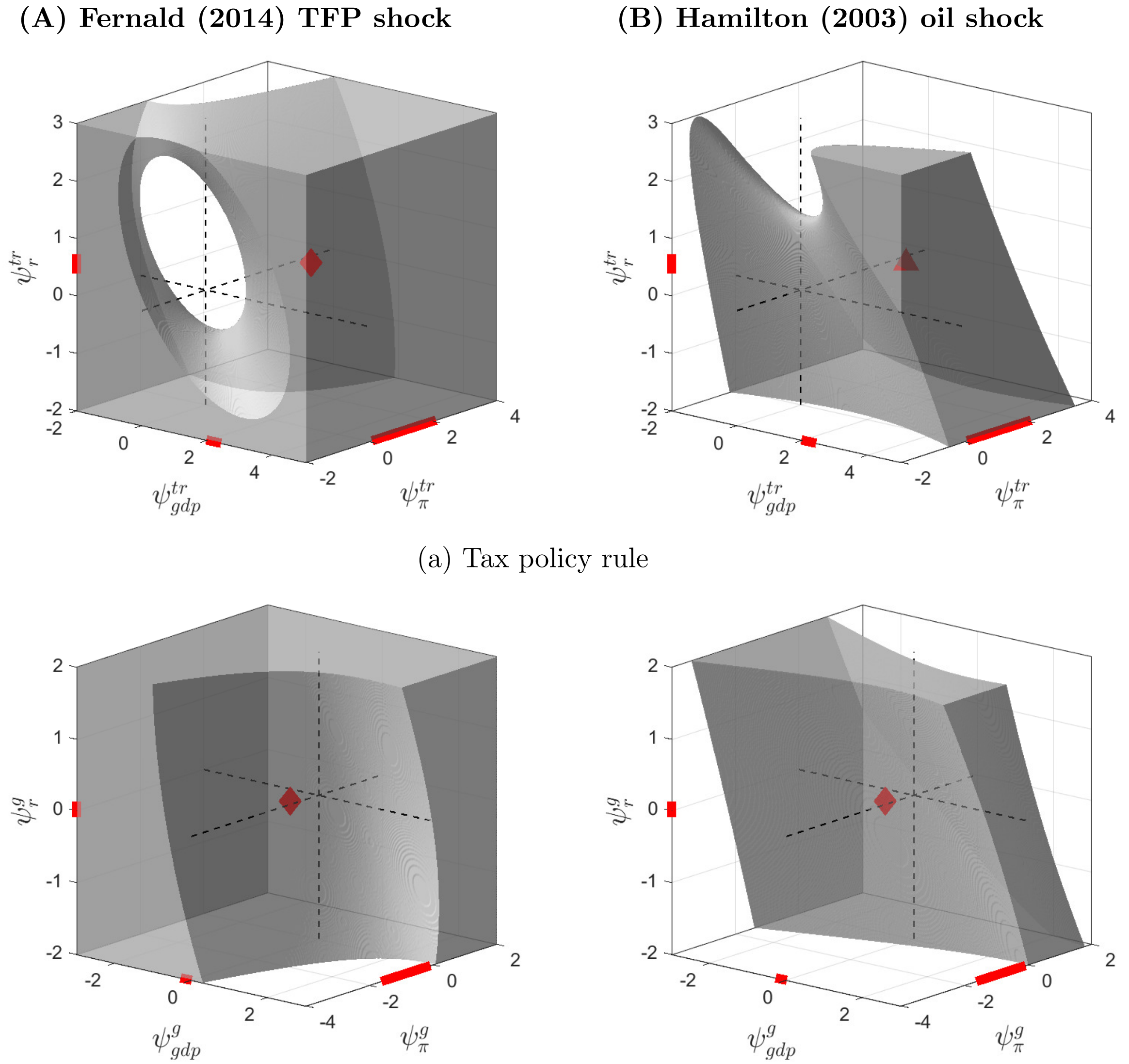
(b) Simple spending rule

Figure E.1: K_a statistic (left axis), and the impact fiscal multiplier (right axis)

Note: This figure presents the K_a statistic (solid line) and the impact multiplier calculated using equation (13) (dashed line) as a function of the systematic response of fiscal policy to output. The triangle represents the limited-information maximum likelihood estimate. The shaded area around the impact multiplier is the pointwise confidence interval obtained from 10,000 bootstrap samples, while the dotted horizontal line indicates the critical value for K_a, both at the 32% significance level. The circle and error bars represent the point estimate and the 68% credible interval from Caldara and Kamps (2017). The dashed vertical line indicates the value of ψ^p_{gdp}, $p \in \{tr, g\}$, where the impact multiplier is zero. Panels (a) and (b) display the results for the simple tax and spending rules, respectively.

E.2 General fiscal policy rules

Figure E.2 presents the 68% confidence sets for the general tax and spending fiscal rules identified either by the TFP proxy of Fernald (2014) or the oil proxy of Hamilton (2003) using AR_a with the χ^2 critical values. The point estimate of Caldara and Kamps (2017) is presented by the diamond (triangle) in the figure if it is (not) included in the confidence sets, and their credible intervals are depicted on the axes. Similarly to the simple fiscal policy rule, using AR_a provides tighter confidence sets than AR^r_a. Although the quantitative conclusion does not change much between AR_a and AR^r_a, this result confirms that using the size-distorted statistic may produce overly narrow confidence sets in practice.



(a) Tax policy rule

(b) Spending policy rule

Figure E.2: 68% confidence sets for the general fiscal policy rules

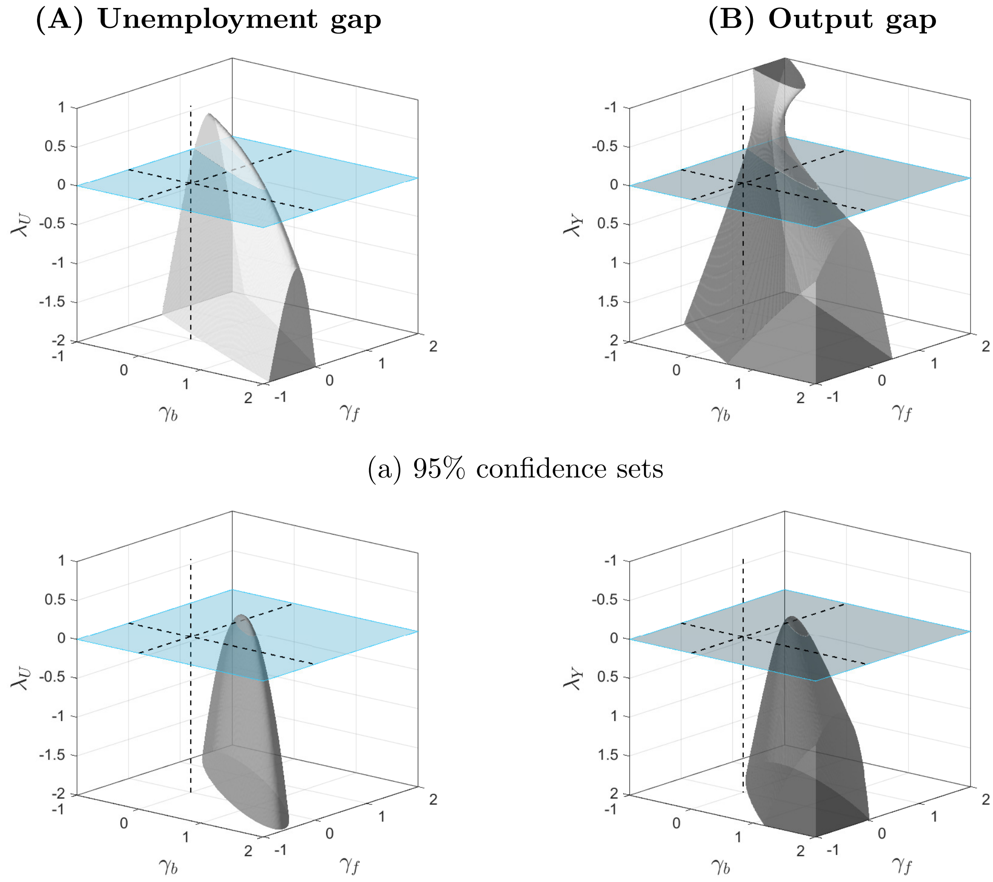
The gray area represents the 68% confidence sets calculated by AR_a with $\chi^2(3)$ critical values. The red lines on the axes show the 68% credible intervals of Caldara and Kamps (2017). The point estimate of Caldara and Kamps (2017) is shown as a red diamond (triangle) if it is (not) included in the confidence set. Panels (a) and (b) correspond to the general tax and spending policy rules, respectively, and columns (A) and (B) show the results identified using the Fernald (2014) TFP proxy and the Hamilton (2003) oil proxy, respectively.

F Robustness checks

This section presents empirical results using AR_a and K_a with the fixed-b critical values to assess the robustness of the results in the main paper that applies AR_a^r or K_a^r with χ^2 critical values.

F.1 New Keynesian Phillips curve

Figures F.1 and F.3 display confidence sets based on the fixed-b critical values of Sun (2014) for the Phillips curve identified by the Romer-Romer proxy (1969—2007) and the HFI proxy (1990–2017), respectively. Figures F.2 and F.4 shows the results under the common restriction $\gamma_b + \gamma_f = 1$. In all figures, the left and right columns correspond to the specifications with the unemployment gap and the output gap as the forcing variables, respectively. The confidence sets are similar to the baseline results presented in the main paper, confirming the robustness of the analysis.

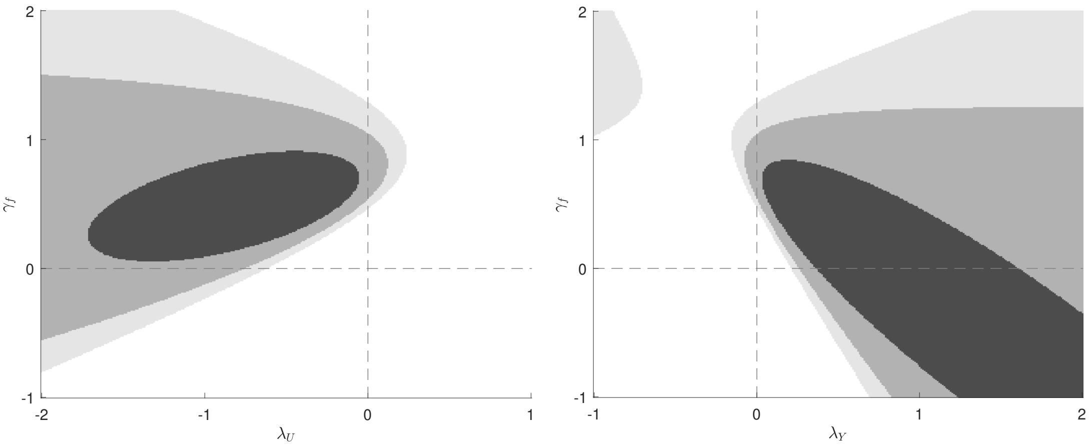


(a) 95% confidence sets

(b) 90% confidence sets

Figure F.1: Phillips curve identified by the Romer-Romer proxy with AR_a and fixed-b cv

Note: Panels (a) and (b) plot the 95 and 90% confidence sets over three parameters $\delta = [\lambda_U(\lambda_Y), \gamma_b, \gamma_f]$, respectively, where the gray areas are the sets that are not rejected by AR_a using the fixed-b critical values of Sun (2014). The left and right columns use the unemployment and output gaps as the forcing variables, respectively. The plane represents the area of a completely flat Phillips curve: $\lambda_U = \lambda_Y = 0$.

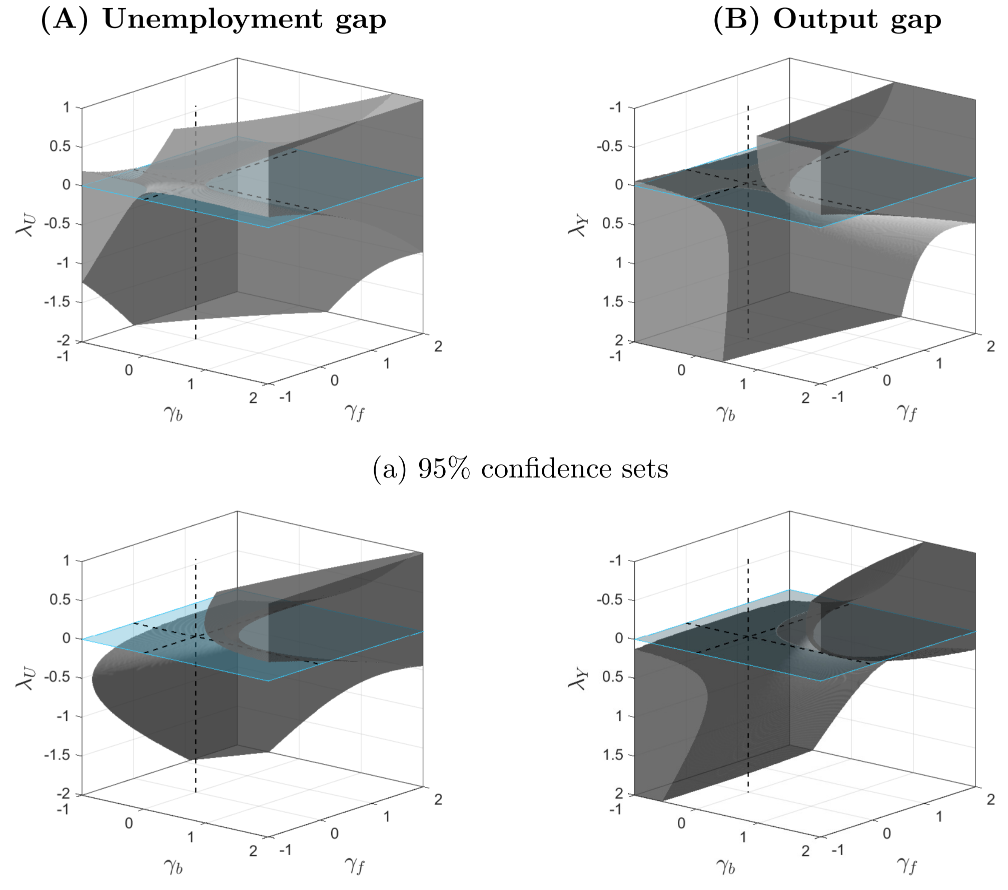


(a) Unemployment gap

(b) Output gap

Figure F.2: With $\gamma_b + \gamma_f = 1$, identified by the Romer-Romer proxy, AR_a and fixed-b cv

Note: This figure shows the 95% (light gray), 90% (gray), and 68% (black) confidence sets for the two parameters γ_f and λ_U (left) or λ_Y (right), under the common restriction $\gamma_b + \gamma_f = 1$.

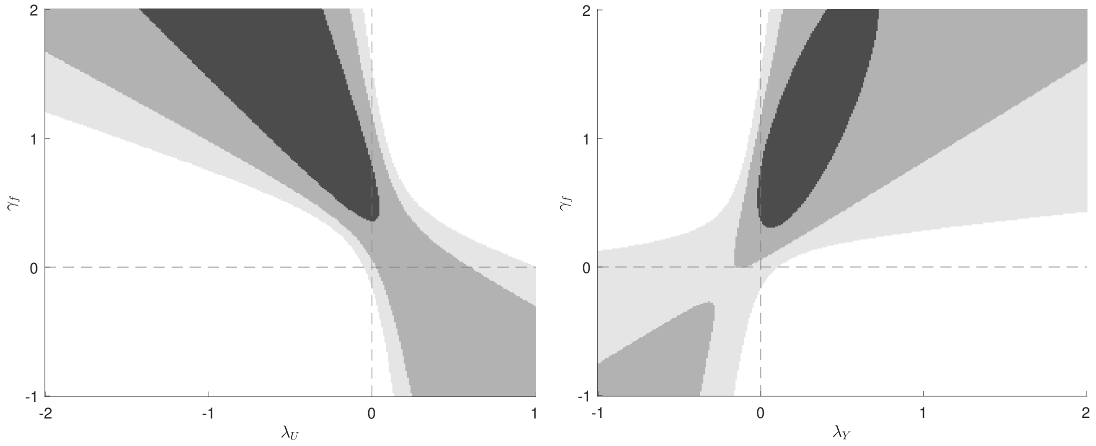


(a) 95% confidence sets

(b) 90% confidence sets

Figure F.3: Phillips curve identified by the HFI proxy with AR_a and fixed-b cv

Note: Panels (a) and (b) plot the 95 and 90% confidence sets over three parameters $\delta = [\lambda_U(\lambda_Y), \gamma_b, \gamma_f]$, respectively, where the gray areas are the sets that are not rejected by AR_a using the fixed-b critical values of Sun (2014). The left and right columns use the unemployment and output gaps as the forcing variables, respectively. The plane represents the area of a completely flat Phillips curve: $\lambda_U = \lambda_Y = 0$.



(a) Unemployment gap

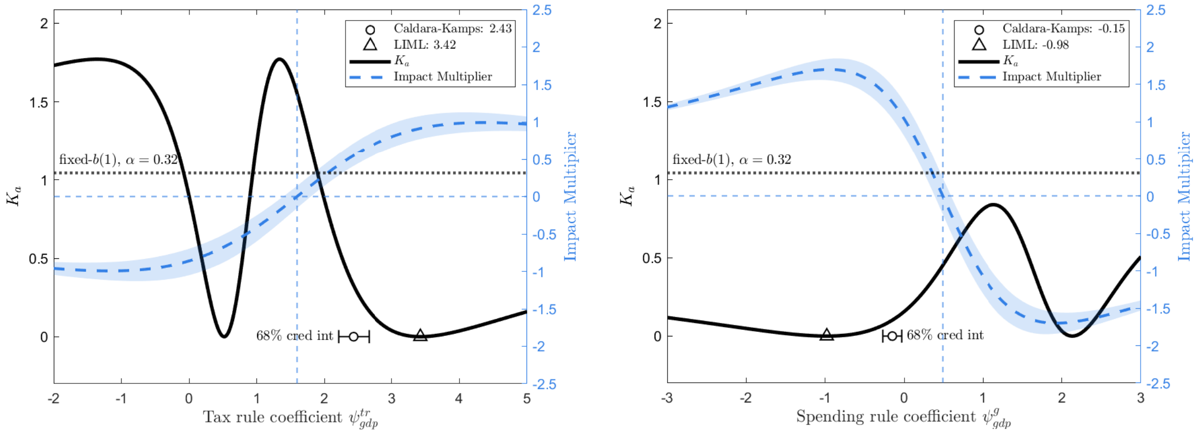
(b) Output gap

Figure F.4: With $\gamma_b + \gamma_f = 1$, identified by the HFI proxy, AR_a and fixed-b cv

Note: This figure shows the 95% (light gray), 90% (gray), and 68% (black) confidence sets for the two parameters γ_f and λ_U (left) or λ_Y (right), under the common restriction $\gamma_b + \gamma_f = 1$.

F.2 Fiscal policy rules and multipliers

Figure F.5 presents the K_a statistic (solid line) and the impact multiplier (dashed line) calculated by equation (13) as a function of the systematic response of fiscal policy to output for the simple tax (ψ^{tr}_{gdp}) and spending (ψ^{g}_{gdp}) rules in equation (12). The dashed vertical line indicates the values of the systematic response for which the impact multiplier is zero. The shaded area around the impact multiplier is the pointwise confidence interval obtained from 10,000 bootstrap samples, while the dotted horizontal line represents the critical value for K_a based on the fixed-b asymptotics of Sun (2014). Following Caldara and Kamps (2017), I adopt a 32% significance level. The results are similar to the findings in the main paper.



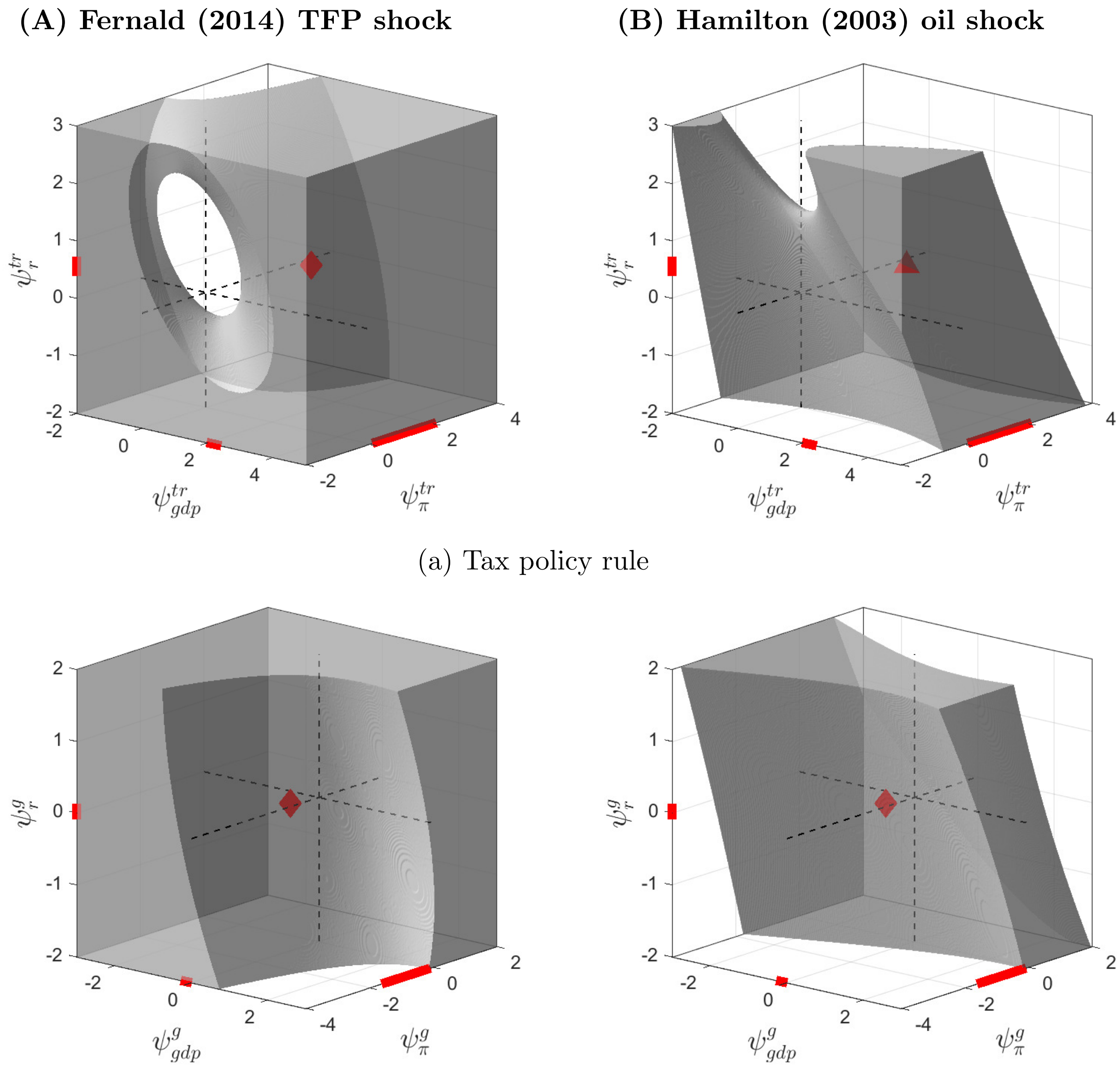
(a) Simple tax rule

(b) Simple spending rule

Figure F.5: K_a statistic (left axis), and the impact fiscal multiplier (right axis)

Note: This figure presents the K_a statistic (solid line) and the impact multiplier calculated using equation (13) (dashed line) as a function of the systematic response of fiscal policy to output. The triangle represents the limited-information maximum likelihood estimate. The shaded area around the impact multiplier is the pointwise confidence interval obtained from 10,000 bootstrap samples, while the dotted horizontal line indicates the critical value based on the fixed-b asymptotics of Sun (2014) for K_a, both at the 32% significance level. The circle and error bars represent the point estimate and the 68% credible interval from Caldara and Kamps (2017). The dashed vertical line indicates the value of ψ^{p}_{gdp}, $p \in \{tr, g\}$, where the impact multiplier is zero. Panels (a) and (b) display the results for the simple tax and spending rules, respectively.

Figure F.6 presents the 68% confidence sets computed using AR_a with fixed-b critical values for the general tax and spending rules (14), identified by either the Fernald (2014) TFP proxy or the Hamilton (2003) oil proxy. The point estimate of Caldara and Kamps (2017) is represented by the diamond (triangle) in the figure if it is (is not) included in the confidence sets, and their credible intervals are depicted on the axes. These results also mostly confirm the robustness of the findings in the main paper. One difference is that the point estimate of Caldara and Kamps (2017) falls outside the confidence set for the tax policy rule when the oil proxy is used. Nevertheless, the robust confidence set remains broadly consistent with the credible set of Caldara and Kamps (2017).



(a) Tax policy rule

(b) Spending policy rule

Figure F.6: 68%confidence sets for the general fiscal policy rules, fixed-b critical values

The gray area represents the 68% confidence sets calculated by AR_a with the fixed-b critical values of Sun (2014). The red lines on the axes show the 68% credible intervals of Caldara and Kamps (2017). The point estimate of Caldara and Kamps (2017) is shown as a red diamond (triangle) if it is (not) included in the confidence set. Panels (a) and (b) correspond to the general tax and spending policy rules, respectively, and columns (A) and (B) show the results identified using the Fernald (2014) TFP proxy and the Hamilton (2003) oil proxy, respectively.