Inference for Macroeconomic Policy Counterfactuals

Daniel J. Lewis*
University College London
& CEPR

Utso Pal Mustafi
Frankfurt School
of Finance & Mangagement

March 5, 2026

Abstract

Recent innovations (McKay and Wolf (2023), Barnichon and Mesters (2023)) have generated considerable interest in policy counterfactuals based on macroeconometric tools. Computation of these counterfactuals constitutes a "regression in impulse response space" that resembles an instrumental variables regression. We use this platform to develop a suite of tools for frequentist inference about policy counterfactuals, including a bias-based for weak identification, a test for whether a counterfactual can be implemented in the data, and methods for inference on the counterfactual responses under both weak and strong identification. To confront the challenge of weak identification, we provide a disciplined approach to select horizons that contain as much information as possible for the counterfactual of interest. These tools provide insights to help applied researchers design better specifications for counterfactual analysis. Only after applying the lessons we draw from our theoretical development can the empirical findings of McKay and Wolf (2023) be confirmed using inference methods robust to weak identification.

JEL classification: C32, E32, E61.
Keywords: Lucas critique, policy counterfactuals, impulse response space, weak identification, robust inference, policy shocks

*daniel.j.lewis@ucl.ac.uk Funded by the European Union (ERC, CreMac, StG-101161596). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We are grateful to Giuseppe Cavaliere as discussant and to Karel Mertens for helpful conversations.

1 Introduction

Identification and estimation of the responses of macroeconomic variables under counterfactual policies is a major challenge in macroeconomics. Lucas (1976) famously highlights how agents' behaviour cannot be expected to remain unchanged following a change in policy parameters; therefore, estimates based on data from an observed prevailing regime are unlikely to be informative for how variables would respond under a counterfactual regime. One strategy to overcome this obstacle is to specify a full structural model (Lucas 1976), while Sims and Zha (2006) offer an early attempt at a reduced form macroeconometric solution. Recently, the estimation of counterfactuals has drawn considerable attention following McKay and Wolf (2023) and Barnichon and Mesters (2023). They show that under relatively minimal structural assumptions – but potentially quite demanding data requirements – the paths of macroeconomic variables under counterfactual policies can be identified from the impulse response functions (IRFs) of those variables to identified shocks. For instance, it is possible to recover how the economy would have responded to a technology shock if the central bank had responded more or less strongly to fluctuations in relevant variables. The influence of these recent papers has already been substantial.

In these new methodologies, the counterfactual responses of macroeconomic variables are obtained as linear combinations of the IRFs of the outcome variable to instruments for the non-policy and policy shocks. To recover the true counterfactual path, the econometrician would require policy news shocks at all horizons. In practice, she must make do with at best a handful, which are used to *approximate* the counterfactual responses. In this paper, we argue that this approximation problem constitutes a "regression in impulse response space" – that is, a least-squares problem where the "observations" are horizons rather than time periods. In particular, the weights used to take the linear combination of IRFs that constitutes the counterfactual are the regression coefficients. Similar problems have recently been studied by Barnichon and Mesters (2020) and Lewis and Mertens (2025b).

Framing the computation of counterfactual paths as relying on a regression in impulse response space highlights previously unrecognised challenges to the methodology. In particular, the denominator of the OLS estimator is the inner product of the IRFs of the counterfactual policy rule equation to the empirical shock measures. Much as in Lewis and Mertens (2025b), this denominator may be statistically close to zero, or at least rank-deficient, when the impulse responses are near-zero, imprecisely estimated, or collinear, the latter of which could be reasonably expected when multiple (correlated) measures of the same policy shock are employed (for instance, Romer and Romer (2004) and Gertler and Karadi (2015) shocks). When this happens, the estimated weights are weakly identified, with finite sample properties reminiscent of a classic weak instruments problem.

To illustrate this problem in practice, we generate data from a 3-equation New

Figure 1: Distribution of counterfactual estimates, 3-equation New Keynesian model

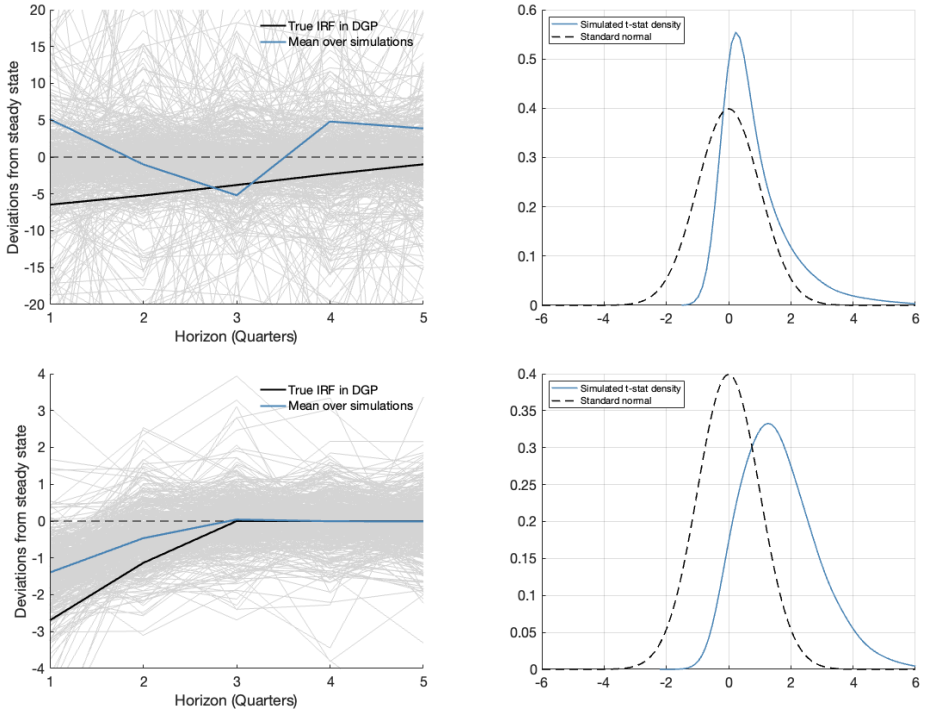


Figure 2: *Notes:*

Notes: Estimated counterfactual output responses to a unit positive cost-push shock in samples simulated ($T = 150$) from a simple New Keynesian model as described in Section 4.5. The top panel considers a model where there are five news shocks available to the econometrician and the bottom a model where there are only two news shocks available, but they are sufficient to implement the counterfactual. In both cases, all five horizons are used to estimate the counterfactual. The left panels plot 500 SVAR estimates from a VAR(4) including the cost-push shock, either five or two monetary policy shocks, output, inflation and interest rate; the mean is computed over 5,000 samples. The right panels plot the estimated density of the $t-$statistic (5,000 samples) for the estimated counterfactual on impact against the standard normal reference distribution.

Keynesian (NK) model, largely calibrated following Galí (2015), create realistic "instruments" by adding noise to the true shocks, estimate IRFs to these instruments using OLS, and use them to estimate the output response to a cost-push shock under a counterfactually more hawkish Taylor Rule, as in McKay and Wolf's (2023) running example (see Section 4.5 for further details on the calibration). Figure 2 displays the distribution of the counterfactual estimates. The top panel considers a case in which the econometrician has five news shocks available and considers five horizons, while the bottom panel considers a case in which she has only two news shocks available, but those two shocks are sufficient to implement the counterfactual rule. The left panels plots the mean estimated counterfactual path over 5,000 draws with $T = 150$ observations each (matching our empirical application), along with the true counterfactual path and 500 estimated counterfactual paths (reduced from 5,000 for readability). In both cases, the counterfactual paths are clearly biased towards zero, particularly at shorter horizons. The right panel plots

the histogram of t-statistics for the estimated counterfactual response on impact. In both cases, the distribution exhibits a strongly skewed right tail, a hallmark of the classic weak instruments problem (see, e.g., Nelson and Startz (1990)). The non-standard distribution also presages that the Gaussian approximation underlying standard inference techniques, (black dashed line) is unlikely to be reliable. Indeed, a conventional 95% t-statistic based confidence interval includes the true counterfactual only 87% (top) and 66% (bottom) of the time, a substantial undercoverage. This pathological behaviour is not confined to frequentist methods. The Bayesian 95% confidence bands used by McKay and Wolf (2023) include the true counterfactual in 99% (top) and 34% (bottom) of samples; taking T to 5,000, the coverages are 99% and 91%. This indicates that the properties of the Bayesian bands are highly sensitive to the DGP, and even in very large samples, do not attain their nominal levels.[1] While Bayesian procedures are not guaranteed to cover the true parameter in finite samples, researchers often assume that they will attain frequentist validity asymptotically (i.e. a Bernstein-von Mises result); here, very large samples indeed are required for that to be the case. Highlighting the features potentially giving rise to weak identification of counterfactuals and demonstrating their consequences is our first contribution.

We next develop a comprehensive framework for frequentist inference for macroeconomic counterfactuals computed following McKay and Wolf (2023). After framing the problem as a regression in impulse response space, the frequentist version of their empirical procedure is a minimum distance estimator. We show that this estimator can be written as a two-stage least-squares (2SLS) estimator and use this result to develop a pre-test for weak identification based on the bias of the weights used to compute the counterfactual paths, a useful diagnostic for the empirical researcher. There is no analogue to such a test in a Bayesian framework, so such a tool is only possible through our frequentist development. This test is our second contribution. Our formulation of the problem makes clear that counterfactuals are likely to be weakly identified and that inference will generally need to account for this fact. If identification is strong, inference for the weights and the counterfactual responses themselves is straightforward. Under weak identification, inference on the weights can be easilyconducted using the Anderson and Rubin (1949) statistic or Kleibergen (2005) statistic, although there can be some subtlety in extending methods to the counterfactual responses. This suite of inference procedures is our third contribution.

One of the challenges highlighted by McKay and Wolf (2023) is that the researcher does not know whether the limited number of shocks she has available contains enough independent information to yield a counterfactual that is a good approximation to the true counterfactual – which might require news shocks at all possible horizons to implement. As a diagnostic, they report responses under

[1]For comparison, t-statistic inference yields 96% and 91%, respectively, in these samples.

perfect implementation of a policy rule for some counterfactuals. We show that it is possible to formally test the null hypothesis that the available shocks contain adequate information to approximate the true counterfactual path through a simple test of the over-identifying restrictions – here, the range of horizons used to estimate the weights, which typically exceeds the number of available shocks. Much like our test for weak identification, this test only arises from the frequentist framework that we develop. This test for the "implementability" of a counterfactual rule is our next contribution.

Our inference procedures demonstrate that certain types of counterfactual paths do not make sense as objects for asymptotic inference. In particular, when a counterfactual rule depends on a single variable – for instance, output gap targeting or an interest rate peg – the confidence band for that variable's counterfactual response must collapse to a *line* at zero *unless the rule is not implementable*. In finite samples, if the corresponding band does not contain zero, this simply means that the shocks available to the econometrician are not adequate to approximate the counterfactual rule. Either way, plotting such a confidence band does not carry economic meaning.

Next, our inference procedures, especially those robust to weak identification, demonstrate that the strength of identification depends starkly on the horizons used for identification. While McKay and Wolf (2023) use 100 horizons in their empirical analysis, doing so in a realistic frequentist VAR is not even possible, where the maximum number of horizons that can be used for inference is at most equal to the number of VAR parameters used to compute the IRFs, but typically much lower. Moreover, even if possible, using additional horizons would not add linearly independent information, up to a second-order approximation. Using horizons for which the estimates are near zero or where the IRFs to the different policy instruments evolve similarly is like adding irrelevant or collinear instruments to an IV model. Such horizons do not add identifying variation, and decrease the aggregate strength of the instruments. The choice of horizons is thus a central decision facing the econometrician. We develop an information criterion to assist with selecting horizons based on a loss function weighing bias (relative to counterfactual responses using *all* feasible horizons) against variance (measured by the volume of a robust confidence set for the weights). Comparing counterfactual responses based on our optimal horizon choice to those using all feasible horizons demonstrates generally negligible changes in the point estimates, but dramatically narrower confidence intervals, which otherwise would generally be *unbounded*. In contrast, with our approach, most of the empirical conclusions of McKay and Wolf (2023) can be replicated using frequentist methods that are robust to weak identification.

The range of issues uncovered and addressed by our frequentist framework means that it should provide useful guidance to applied econometricians seeking to apply the McKay and Wolf (2023) identification result. However, one might

argue that while such observations are interesting, in practice, inference can still just proceed using McKay and Wolf's (2023) Bayesian inference procedure. We contend that this approach is unsatisfactory for two reasons. First, for a principled Bayesian, the procedure is not a formal Bayesian treatment. Inference for the counterfactuals is indirect, with posterior distributions constructed for the underlying VAR IRFs, and the implied distribution backed out for the impulse responds weights and then the counterfactual responses. This does not allow the econometrician to place a prior on the object of interest – the counterfactual responses – or even the weights, which hinders the transparency of the procedure. A common Bayesian rebuttal to weak identification is that, provided the prior reflects the researcher's true beliefs, the associated posterior is always valid, so there is no cause for concern. However, priors in McKay and Wolf (2023) are applied to the VAR parameters, which are known to be typically consistently estimated, not the weakly identified weights or counterfactual responses, so, in the case of weak identification, when the data offers limited information through the likelihood or the influence of the prior may not vanish, the implications of the priors for the parameters of interest are largely opaque and unlikely to accord with the econometrician's intentions. Second, for many applied researchers, Bayesian inference methods are often adopted for practical purposes, motivated by the fact that they have frequentist guarantees asymptotically, following a Bernstein-von Mises argument. Notably, such guarantees do not generally hold in the presence of misspecification (for instance where a T-period counterfactual rule is approximated using a small number of a available policy shock instruments), see for instance Kleijn and van der Vaart (2012). Second, Bernstein-von Mises results for *functions of model parameters* (like counterfactual responses in this setting) hinge on regularity conditions that hold only under strong identification. To the extent that any equivalence continues to hold approximately, Bayesian inference would approximate *invalid* frequentist procedures (i.e. those based on standard maximum likelihood asymptotics), as the results noted above for our simple illustration reflect. Related phenomena have recently been documented for the IV regression model by Giannone et al. (2025). Therefore, whether for a purist or practical Bayesian, there is reason to be dissatisfied with existing inference options.

The paper proceeds as follows. Section 2 introduces the problem and provides a step-by-step guide for empirical researchers seeking to implement our methods. Section 3 presents the McKay and Wolf (2023) problem as a minimum distance estimator and develops tests for weak identification. Section 4 develops inference methods and tests for implementability of the counterfactual rule. We then discuss the problem of selecting horizons for estimation and propose an information criterion do so, demonstrating the performance of our methods in model-generated data. Section 4.4 Section 5 applies our methods to the empirical counterfactuals studied by McKay and Wolf (2023), illuminating how our recommendations can

guide the estimation of counterfactual responses in practice and lead to economically meaningful findings.

2 Overview and guide for applied users

In this section we sketch the setting, explain the challenges to idenfication, and provide a step-by-step guide for applied users.

2.1 Setting

The key insight of McKay and Wolf (2023) is an identification result that allows researchers to conduct inference on the behaviour of variables under counterfactual policy rules using only widely-available estimates of the IRFs of macro aggregates to empirical measures of structural shocks. To study how variables respond to a non-policy shock under a counterfactual policy rule, Sims and Zha (2006) subject the economy to a new policy shock at each time t such that the counterfactual rule holds, t-by-t. However, the Lucas (1976) critique suggests that is not enough: to fully match the behaviour of the economy under the counterfactual rule, it must be expected to hold in the current period *and all future periods*. McKay and Wolf (2023) show that under fairly general assumptions, a sequence of *policy news shocks* to the prevailing policy rule at all future horizons and knowledge of IRFs of relevant variables to those shocks under the prevailing rule is all that is necessary to enforce the counterfactual policy rule and subsequently recover the behaviour of those variables under that rule.

We begin by sketching a version of the McKay and Wolf (2023) setup and identification argument. For a detailed development, the interested reader should consult the original paper. They assume that the economy is defined by n_x observable endogenous variables x_t, n_z policy variables z_t, and n_ε non-policy shocks ε_t. The values of the endogenous variables are implicitly defined by a system of Tn_x linear equations. For example, in a simple three-equation NK model, $x_t \equiv (y_t,\ \pi_t)'$ and these equations are the Euler equation and Phillips curve:

$$y_t - y_{t+1} - \frac{1}{\gamma}\pi_{t+1} + \frac{1}{\gamma}i_t = 0 \tag{1}$$

$$\pi_t - \kappa y_t - \beta\pi_{t+1} - \varepsilon_t - \theta\varepsilon_{t-1} = 0. \tag{2}$$

The values of the policy variables are defined by the policy rule

$$A_x\mathbf{x} + A_z\mathbf{z} + \boldsymbol{\nu} = \mathbf{0}, \tag{3}$$

a system of Tn_z equations, where ν_t is a ν_z-vector of policy shocks, and boldface vectors $\{x, z, \varepsilon, \nu\}$ stack time paths for all variables (e.g., $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_{n_x}')'$). Typically, $n_z = 1$, for instance, z_t is the nominal interest rate. As a simple example,

consider a Taylor Rule, as in McKay and Wolf (2023),

$$i_t - \phi\pi_t - \nu_{0,t} - \nu_{1,t-1} = 0 \tag{4}$$

where $\nu_{0,t}$ is a contemporaneous policy shock and $\nu_{1,t-1}$ is the one-period ahead news shock realised at $t-1$. Then $z_t \equiv i_t$ and

$$A_x = \begin{bmatrix} \mathbf{0}_{T\times T} & -\phi I_T \end{bmatrix}, \; A_z = I_T. \tag{5}$$

Note that this structure is fully general in the sense that it allows the value of the policy variables in any time period to depend on the values of variables (and thus shocks) in any other time period since A_x is $(Tn_z)\times(Tn_x)$ and A_z is $(Tn_z)\times(Tn_z)$.

The econometrician wants to study the behaviour of the economy under some counterfactual policy rule,

$$\tilde{A}_x\mathbf{x} + \tilde{A}_z\mathbf{z} = \mathbf{0}. \tag{6}$$

For instance, in the Taylor Rule example above, $i_t = \tilde{\phi}\pi_t$, $\tilde{A}_x = \begin{bmatrix} \mathbf{0}_{T\times T} & -\tilde{\phi} I_T \end{bmatrix}$, with $\tilde{A}_z = A_z$.

The econometrician seeks to recover how the economy would have responded to a scalar non-policy shock ϵ_t at time 0 (and zero thereafter), where $\epsilon \in \varepsilon$. Due to the linearity of the economy, the paths of all variables in response to shocks can be fully described by IRFs, i.e. $\mathbf{x}_A(\epsilon_0) = \Theta_{x,\epsilon,A}$ where $\Theta_{x,\epsilon,A}\epsilon_0$ is a Tn_x matrix stacking the $T-$horizon IRFs of $x_{1t}, \ldots, x_{n_x t}$.

McKay and Wolf's (2023) approach is to find a value for the policy shock vector of contemporaneous and news shocks $\boldsymbol{\nu}$ such that, following ϵ_0, the counterfactual rule (6) holds exactly. The idea is that the right combination of policy news shocks can convey that the policy variable z_t will follow the exact same path dictated by the rule. The vector of news shocks imposing the rule, denoted $\tilde{\boldsymbol{\nu}}$, can be obtained by plugging the responses of the variables into the counterfactual policy rule and solving,

$$\tilde{A}_x\left[\Theta_{x,\epsilon,A}\epsilon_0 + \Theta_{x,\nu,A}\times\tilde{\boldsymbol{\nu}}\right] + \tilde{A}_z\left[\Theta_{z,\epsilon,A}\epsilon_0 + \Theta_{z,\nu,A}\times\tilde{\boldsymbol{\nu}}\right] = \mathbf{0}, \tag{7}$$

Essentially, $\tilde{\boldsymbol{\nu}}$ is the date-0 policy news shock vector that mimics the equilibrium path of z_t under the counterfactual rule; the first term in square brackets is the response of $\mathbf{x}$ under the counterfactual rule, and the second term is the response of $\mathbf{z}$ under the counterfactual rule.Then, the counterfactual responses are

$$\mathbf{x}_{\tilde{A}}(\epsilon_0) = \Theta_{x,\epsilon,\tilde{A}}\epsilon_0 = \Theta_{x,\epsilon,A}\epsilon_0 + \Theta_{x,\nu,A}\times\tilde{\boldsymbol{\nu}} \tag{8}$$

$$\mathbf{z}_{\tilde{A}}(\epsilon_0) = \Theta_{z,\epsilon,\tilde{A}}\epsilon_0 = \Theta_{z,\epsilon,A}\epsilon_0 + \Theta_{z,\nu,A}\times\tilde{\boldsymbol{\nu}}. \tag{9}$$

Of course, the researcher will never have access to an estimate of policy news shocks at all future horizons, and thus cannot estimate the set of IRFs at horizons

from 0 to $T-1$, $\Theta_{i,\nu,A}$, which is required to implement (6) at all time periods. Instead, McKay and Wolf (2023) propose to *approximate* the rule using a lower-dimensional set of K empirical policy instruments, m_t, that are assumed to be a linear combination of the idealised policy news shocks, ν_t, and some smaller set of horizons, say $0, 1, \ldots, H-1$. Suppose an instrument is also available for ϵ_t, e_t. Denote the prevailing IRFs to e_t for horizons $0, 1, \ldots, H-1$, as $\Pi_{i,A}$ and those to m_t as $\Omega_{i,A}$. While $H < T$, H can be chosen such that for $h > H$, the effect of a time-t shock has fully dissipated, so there is no information lost by omitting longer horizons (McKay and Wolf (2023) set $H = 100$). Let $\tilde{A}_x^H$ select the first H rows of $\tilde{A}_x$ corresponding to each policy variable and the first H columns corresponding to each endogenous variable.[2] Then the econometrician aims to recover

$$\beta = \text{argmin}_b \left|\left|\tilde{A}_x^H \left(\Pi_{x,A} + \Omega_{x,A} \times b\right) + \tilde{A}_z^H \left(\Pi_{z,A} + \Omega_{z,A} \times b\right)\right|\right| \tag{10}$$

$$= \text{argmin}_b \left|\left|\Psi + \Phi b\right|\right|, \tag{11}$$

where

$$\Psi \equiv \tilde{A}_x^H \Pi_{x,A} + \tilde{A}_z^H \Pi_{z,A} \tag{12}$$

$$\Phi \equiv \tilde{A}_x^H \Omega_{x,A} + \tilde{A}_z^H \Omega_{z,A}, \tag{13}$$

This minimisation problem yields

$$\beta = -\left(\Phi'\Phi\right)^{-1} \Phi'\Psi. \tag{14}$$

Then, the approximate counterfactual response of variable c is simply

$$\Pi_{c,\tilde{A}} = \Pi_{c,A} + \Omega_{c,A} \times \beta \tag{15}$$

In practice, the econometrican can plug in $\hat{\Pi}_{c,A}$ and $\hat{\Omega}_{c,A}$ (using a VAR or local projection) for $\Pi_{c,A}$ and $\Omega_{c,A}$ to obtain $\hat{\beta}$ and $\hat{\Pi}_{c,\tilde{A}}$.

Equation (14) makes clear that obtaining the weights, β, used to construct the counterfactual (15) is a *regression in impulse response space*. That is, it is an OLS regression in which the response of the counterfactual policy rule to non-policy instrument e_t is projected onto the response of the counterfactual policy rule to policy instruments m_t. (11) is then just like any other OLS regression problem, except that all of the variables are generated regressors (estimated IRF coefficients). As a result, β is subject to any challenges faced by such projection estimators.

In particular, we know that the behaviour of $\hat{\beta}$ may be non-standard when estimation error means that the denominator of the OLS estimator is rank deficient, or at least not statistically far from rank deficient in finite samples. When this

[2] Concretely, $\tilde{A}_x^H = \left(\begin{bmatrix} I_H & \mathbf{0}_{H\times(T-H)} \end{bmatrix} \otimes I_{n_z}\right) \tilde{A}_x \left(\begin{bmatrix} I_H & \mathbf{0}_{H\times(T-H)} \end{bmatrix} \otimes I_{n_x}\right)'$ and similarly for $\tilde{A}_z^H$ (i.e. $\tilde{A}_z^H = \left(\begin{bmatrix} I_H & \mathbf{0}_{H\times(T-H)} \end{bmatrix} \otimes I_{n_z}\right) \tilde{A}_z \left(\begin{bmatrix} I_H & \mathbf{0}_{H\times(T-H)} \end{bmatrix} \otimes I_{n_z}\right)'$.

happens, the estimator is confounded by what is essentially a division by zero problem. This is at the core of the weak instruments literature (see, e.g., Nelson and Startz (1990), Stock and Yogo (2005)) and drives the non-standard behaviour of the counterfactual estimates in Figure 2, discussed in the introduction. Such weak identification problems bias $\hat{\beta}$ (and thus $\hat{\Pi}_{i,\tilde{A}}$) and mean that the standard normal approximations typically relied upon for inference likely provide a poor approximation of finite sample behaviour, as in the discussion of Figure 2.

2.2 Sources of identification breakdown

In practice, the denomninator of $\hat{\beta}$ can be near-zero in a statistical sense for a variety of reasons. First, the policy shock measures may have only a modest impact on the variables entering the policy rule (inflation and the nominal interest rate in the Taylor Rule example). It has been well-documented that many modern high-frequency measures of monetary policy shocks consist of quite small changes in financial market expectations around FOMC announcements, and, as a result, explain relatively little variation in inflation (e.g., Ramey (2016)), although they may still explain considerable variation in the nominal interest rate. If the estimated responses are small, then sampling variation may render the denominator close to zero in finite samples. Next, even if the true responses are substantially non-zero, there is often considerable statistical uncertainty around IRF estimates, particularly at longer horizons, and especially for local projections. In this case, the sampling distribution of the denominator includes values close to zero with non-trivial probability. Third, a key challenge of the McKay and Wolf (2023) methodology is finding linearly independent measures of policy shocks. The researcher will typically hope that two measures of monetary policy shocks contain different linear combinations of news about the current and future policy rate, for instance. McKay and Wolf (2023) consider two pairs of monetary policy shocks, Romer and Romer (2004) plus Gertler and Karadi (2015) and Aruoba and Drechsel (2024) plus Miranda-Agrippino and Ricco (2021). To the extent that in principle each of these shocks is trying to measure the same thing (news about monetary policy released on a given day), it is reasonable to wonder whether, in population, the impulse responses to any two measures are linearly independent (and if they are not, whether the only reason measurement error); if two instruments measure the same monetary policy shock, the IRFs should be collinear. In such cases, the denominator would be close to rank one, even if the IRFs themselves are distinct from zero. Finally, in a stationary economy, variables should return to their steady state values in the long run, so IRFs should return to zero for large H; McKay and Wolf (2023) use $\mathcal{H} = \{0, 1, \ldots, 100\}$. Appendix D shows that identification strength depends on the average signal over the horizons considered, so including many zero responses washes out identification from strongly responsive horizons.

2.3 Guide for applied practice

The focus of this paper is to highlight these challenges and confront them head-on. We provide a step-by-step guide for applied use below. First, a novel information criterion, tailored to weak identification, assists the researcher in selecting horizons to use for estimation. Next, the econometrician can test whether the weights required to implement a counterfactual rule are weakly identified. Such tests can also be used to compare identification strength across different choices of horizons and instruments. After obtaining estimated weights, $\hat{\beta}$, a test for the "implementability" of a counterfactual tells the researcher whether the available instruments are adequate to plausibly identify behaviour under a counterfactual rule. Finally, we explain how inference can be conducted of the counterfactuals of interest. In what follows, we assume that

$$\sqrt{T}\left(\begin{bmatrix}\text{vec}(\hat{\Pi}_{x,A})\\ \text{vec}(\hat{\Pi}_{z,A})\\ \text{vec}(\hat{\Omega}_{x,A})\\ \text{vec}(\hat{\Omega}_{z,A})\end{bmatrix} - \begin{bmatrix}\text{vec}(\Pi_{x,A})\\ \text{vec}(\Pi_{z,A})\\ \text{vec}(\Omega_{x,A})\\ \text{vec}(\Omega_{z,A})\end{bmatrix}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{M}), \tag{16}$$

as in our formal development below. Let $\mathbf{M}_2$ denote the block of $\mathbf{M}$ corresponding to the IRFs to policy shocks, $\Omega_{i,A}$, and $\hat{\mathbf{M}}$ a consistent estimator of $\mathbf{M}$.

1. **Setup and specification selection.**

   (a) Using VARs, local projections, or a similar method, obtain estimates of the impulse responses to the non-policy shock, using an available instrument, e, $\hat{\Pi}_{x,A}$ and $\hat{\Pi}_{z,A}$. Obtain estimates of the impulse responses to the policy shocks, $\hat{\Omega}_{x,A}$, and $\hat{\Omega}_{z,A}$. In the simple three-variable NK example, these are the IRFs of the output gap, y_t, inflation, π_t, and the nominal interest rate, i_t, to a cost push shock and some number of monetary policy shocks, respectively.

   (b) Select the horizons, $\mathcal{H}$, used to estimate the counterfactual rule. In a VAR, $H \equiv |\mathcal{H}|$ cannot exceed $H^{\max} \equiv pn^2 + (K+1)n$, where p is the lag length and n is the number of variables. In general, the maximum feasible number of horizons, denoted $\bar{H}$, will be considerably lower, still, see Remark 1 for details. We propose an information criterion based on a loss function trading off bias (of $\hat{\Pi}_{\tilde{A}}$ compared to using all feasible horizons, $0, \ldots, \bar{H}-1$) and variance (measured by the volume of a robust confidence set for β). To find the optimal set of horizons according to this rule, obtain the set $\mathcal{H}^*$ that minimises the loss function in (55). We place greater weight, $w = 0.75$, on bias for our empirical analysis. Note that counterfactual responses may still be computed for any horizons of interest, not just those used for estimation.

   (c) Conduct a first-stage test for weak identification of β. Construct the test

statistic as

$$g_{\min} = T^{-1} \text{mineval}\{\hat{\Delta}^{-\frac{1}{2}} \hat{\Phi}(\mathcal{H}^*)' \hat{\Phi}(\mathcal{H}^*) \hat{\Delta}^{-\frac{1}{2}}\}, \tag{17}$$

$\hat{\Phi}(\mathcal{H}^*)$ collects the rows of $\hat{\Phi}$ corresponding to the horizons in $\mathcal{H}^*$, and $\hat{\mathbf{W}}_2$ estimates the asymptotic variance of $\text{vec}(\hat{\Phi}(\mathcal{H}^*))$ (see (38) for an explicit formula). Critical values are non-standard, but can be constructed using the supplied Matlab program `cfweakivtest.m`, for a user-supplied bias tolerance, τ, and significance level, α (i.e., 0.10 and 0.05 respectively, as in the Stock and Yogo (2005) rule of thumb).

2. **Estimation and implementability.**

   (a) Estimate $\hat{\beta}$,

$$\hat{\beta} = -\left(\hat{\Phi}(\mathcal{H}^*)' \hat{\Phi}(\mathcal{H}^*)\right)^{-1} \hat{\Phi}(\mathcal{H}^*)' \hat{\Psi}(\mathcal{H}^*), \tag{18}$$

   where $\hat{\Psi}(\mathcal{H}^*)$ collects the rows of $\hat{\Psi}$ corresponding to horizons $\mathcal{H}^*$.

   (b) Conduct a test for implementability of the counterfactual rule. Compute Hansen's J-statistic,

$$\hat{J} = T(\hat{\Psi}(\mathcal{H}^*) - \hat{\Phi}(\mathcal{H}^*)\hat{\beta})' \left(\left(\begin{bmatrix} 1 & -\hat{\beta}' \end{bmatrix} \otimes I_H \right) \hat{\mathbf{W}} \left(\begin{bmatrix} 1 & -\hat{\beta}' \end{bmatrix}' \otimes I_H \right) \right)^{-1} (\hat{\Psi}(\mathcal{H}^*) - \hat{\Phi}(\mathcal{H}^*)\hat{\beta})$$

   where $\hat{\mathbf{W}}$ estimates the asymptotic variance of $\left(\hat{\Psi}(\mathcal{H}^*)', \text{vec}(\hat{\Phi}(\mathcal{H}^*))'\right)$, following the explicit formula in (38)-(39). Under strong identification, compare this to χ^2_{H-K} critical values to test the null hypothesis that the rule is implementable. Under weak identification, use χ^2_H critical values. If the null is rejected, consider alternative instruments or a smaller set of horizons, or a different counterfactual entirely.

   (c) Construct counterfactual(s) for variable(s) i as

$$\hat{\Pi}_{i,\tilde{A}} \equiv \hat{\Pi}_{i,A} + \hat{\Omega}_{i,A} \times \hat{\beta}. \tag{19}$$

3. **Inference.**

   (a) **Strong identification.** Construct a confidence interval for a given horizon using the t-statistic

$$t_i^h = \frac{\hat{\Pi}^h_{i,\tilde{A}}}{\sqrt{V_{\Xi,h}/T}}, \tag{20}$$

   where $\Pi^h_{i,\tilde{A}}$ is the h^{th} entry of $\Pi^h_{i,\tilde{A}}$ and $V_{\Xi,h}$ is the h^{th} diagonal element of the variance matrix in (48).

   (b) **Weak identification.** Construct an Anderson and Rubin (1949) or Kleibergen (2005) confidence interval for $\Pi^h_{i,\tilde{A}}$, using the test statistics and associated critical values in (49) and (50), respectively. For further

details about how to construct these sets in practice, see Remark 7.

3 Estimation

In this section, we present a framework for frequentist estimation of macroeconomic counterfactuals. We begin by outlining key assumptions. We then present $\hat{\beta}$ and $\hat{\Pi}_{i,\tilde{A}}$ as minimum distance estimators. Finally, we develop our test for weak identification.

Assumption 1. *The non-policy instrument e_t is a valid instrument for the non-policy shock ϵ_t and the policy instruments m_t are valid instruments for the column space of contemporaneous and future policy news shocks, ν_t. Therefore, the IRFs to e_t, $\Pi_{i,\tilde{A}}$, are a scalar multiple of the first H horizons of $\Theta_{i,\epsilon,A}$ and the IRFs to m_t, $\Omega_{i,A}$, are a linear combination of the first H horizons of $\Theta_{i,\nu,A}$.*

Assumption 1 is a high-level assumption that relates the IRFs estimable by the econometrician using the available instruments to the IRFs to the true shocks.

It will be convenient to define the Tn_z vector

$$\zeta = \tilde{A}_x^H \mathbf{x} + \tilde{A}_z^H \mathbf{z}. \tag{21}$$

That is, ζ_t is equal to the value(s) of the counterfactual policy rule at time t. By linearity, the IRF of ζ_t to e_t is equal to Ψ and those to m_t are equal Φ, both defined above. If the counterfactual rule holds, by definition $\zeta_t = 0, \forall, t = 1, \ldots, T$, so this must also be true conditional on a realisation of a shock, ϵ_t; therefore, the set of policy news shocks that implement the rule (if such a set exists) satisfy

$$\Psi + \Phi\beta = \mathbf{0}_{H\times 1}. \tag{22}$$

This constitutes an $H-$dimensional distance function for estimation,

$$f(b) = \hat{\Psi} - \hat{\Phi}b, \tag{23}$$

with identity-weighted minimum distance objective

$$S(b; I_H) = f(b)' f(b) \tag{24}$$

and solution

$$\hat{\beta} = (\hat{\Phi}'\hat{\Phi})^{-1}\hat{\Phi}'\hat{\Psi}. \tag{25}$$

Expressing $\hat{\beta}$ as a minimum distance estimator will facilitate much of the discussion of inference below. For now, we suppress the dependence of $\hat{\beta}$ on the choice of horizons. Implicitly, (25) depends on the horizons contained in $\hat{\Psi}$ and $\hat{\Phi}$. We make this dependence explicit when we discuss the choice of horizons in Section 4.4 (and in the guide for applied practice above).

Optimal Policy. McKay and Wolf (2023) also consider optimal policy exercises based on a policymaker's loss function, rather than an explicit counterfactual policy rule. Consider the loss function

$$\mathfrak{L} = \frac{1}{2}\sum_{i=1}^{n_x} \lambda_i x_i' \mathcal{W} x_i, \tag{26}$$

where λ_i is the relative weight placed on variable i in the policymaker's objective function, and the weighting matrix $\mathcal{W}$ is the objective function's weighting across horizons. As they explain, the optimal policy can be estimated using particular values for $\tilde{A}_x^H, \tilde{A}_z^H$. Following page 13 of their Supplement, $\tilde{A}_x^H = \Lambda^{\frac{1}{2}} \otimes \mathcal{W}^{\frac{1}{2}}$, and $\tilde{A}_z^H = 0$. With these definitions, β can be estimated as before, and inference can proceed using the same methods described below.

3.1 Testing for weak identification

It is not, in general, feasible to provide bias-based tests for weak identification for minimum distance estimators. However, in this section we derive a representation of (25) as an instrumental variables estimator that permits us to develop such a test. For simplicity, our development focuses on the case of local projections, but we explain below how to apply it to VARs.

Let ζ_H denote the $(Tn_z) \times H$ matrix of current and future values of ζ_t (i.e. $\zeta_{Ht} = [\zeta_t,\ \zeta_{t+1},\ \ldots,\ \zeta_{t+H-1}]$. Additionally, let $\Phi = C/\sqrt{T}$, which, under the assumption that available shock measures are adequate to approximate the desired counterfactual rule, (22), implies that $\Psi = C\beta/\sqrt{T}$. This device models the IRFs to m_t as being local-to-zero, as is conventional in the weak instruments literature (see, e.g., Stock and Yogo (2005)). For the purposes of this section we assume that ζ_H can be written as

$$\zeta_H = e\Psi' + m\Phi' + v_H. \tag{27}$$

This is implied, for instance, if $(x_t', z_t')'$ is generated as a vector moving average (see, e.g., the "impulse propagation paradigm" in Plagborg-Møller and Wolf (2021)) and the non-policy instrument e is uncorrelated with the policy instruments, m. Then

$$\zeta_H = e\beta' C'/\sqrt{T} + mC'\sqrt{T} + v_H. \tag{28}$$

By Frisch-Waugh-Lovell, any controls used in estimating the IRFs of ζ_t (equivalently, x_t and z_t) can be partialed out, allowing (28) to hold without loss of generality, as is standard in the weak instruments literature. Let $\xi_t = [e_t,\ m_t']'$ and ξ be the $T \times (K+1)$ matrix stacking the shock measures over time. We make the following formal assumptions:

Assumption 2.

1. $\Psi = \Phi\beta = C\beta/\sqrt{T}$, *where* C *is fixed and full-rank;*

2. *The empirical shock measures are standardised such that* $\xi'\xi/T = I_{K+1}$;

3. $E[v_{Ht}] = 0$, $E[v_{Ht}\xi_t'] = 0$;

4.
$$\sqrt{T}\begin{pmatrix}\hat{\Psi}\\ \mathrm{vec}(\hat{\Phi})\end{pmatrix} \xrightarrow{d} \begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}C\beta\\ \mathrm{vec}(C)\end{pmatrix}, \mathbf{W}\right), \tag{29}$$

5. $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$.

The first point assumes that the available policy shock measures suffice to implement the chosen counterfactual and the local-to-zero device. The normalisation in 2 is without loss of generality, except for the imposition of orthogonality between e_t and m_t; while this holds by definitionally for the true shocks, it may need to be imposed by transformation for empirical measures. The first part of point 3 will in general follow from 1 and 2; in conjunction with 1, the latter part requires that the empirical shock measures are valid instruments for non-policy and policy shocks, respectively, and follows from Assumption 1. Point 4 is a high-level central limit theorem for the IRF estimates. Part 5 assumes that a consistent (possibly robust) estimator for the asymptotic variance of the IRFs is available. Since $\hat{\Psi} = \zeta_H' e$ and $\hat{\Phi} = \zeta_H' m$,

$$\hat{\beta} = -\left(m'\zeta_H\zeta_H'm\right)^{-1} m'\zeta_H\zeta_H'e. \tag{30}$$

This represents $\hat{\beta}$ as an instrumental variables estimate in the regression of the non-policy instrument, e, on the policy instruments, m, using the policy rule at various leads as instruments, ζ_H. This is far from obvious, since the empirical shock measures become the regressors. By Assumption 2, there is no endogeneity problem to overcome, but the estimator nevertheless has the structure of a 2SLS estimator, where the projection of e on ζ_H is regressed on the projection of m on ζ_H.

Proposition 1. *Under Assumptions 1 and 2,*

$$\hat{\beta} - \beta \xrightarrow{d} \beta^* = \left(R_{K,H}'(\eta_2\eta_2' \otimes I_H)R_{K,H}\right)^{-1} R_{K,H}'\,\mathrm{vec}(\tilde{\eta}_1\eta_2'), \tag{31}$$

where $\tilde{\eta}_1 = \eta_1 - (\beta' \otimes I_H)\eta_2$, $R_{K,H} = I_K \otimes \mathrm{vec}(I_H)$, *and*

$$\begin{pmatrix}\tilde{\eta}_1\\ \eta_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\ \mathrm{vec}(C)\end{pmatrix}, \mathbf{S}\right). \tag{32}$$

Proof. The proposition follows directly from the stated assumptions, (30), and the continuous mapping theorem. □

The matrices $\mathbf{S}$ and $\mathbf{W}$ are paritioned with $\mathbf{S}_1, \mathbf{W}_1$ corresponding to $\hat{\Psi}$ and

$\mathbf{S}_2, \mathbf{W}_2$ corresponding to $\text{vec}(\hat{\Phi})$, with

$$\begin{aligned}\mathbf{S}_1 &= \mathbf{W}_1 + (\beta' \otimes I_H)\mathbf{W}_2(\beta \otimes I_H) - (\beta' \otimes I_H)\mathbf{W}_{12}' - \mathbf{W}_{12}(\beta \otimes I_H), \\ \mathbf{S}_{12} &= \mathbf{W}_{12} - (\beta' \otimes I_H)\mathbf{W}_2 \ , \ \mathbf{S}_2 = \mathbf{W}_2.\end{aligned} \tag{33}$$

Proposition 1 shows that under weak instruments asymptotics, $\hat{\beta}$ does not have a degenerate limit as $T \to \infty$, instead converging in distribution to a function of normal random variables; it is not consistent for β. This distribution matches that developed by Lewis and Mertens (2025a) for 2SLS estimators (letting K in that paper equal H in this and N in that paper equal K in this), since (30) represents $\hat{\beta}$ as a 2SLS estimator.

We thus adapt the testing procedure developed in that paper to the present setting. First, we propose to measure the asymptotic bias of $\hat{\beta}$ as

$$B = \sqrt{\frac{E\left[\beta^*\right]' \Omega_{x_i,A}' \Omega_{x_i,A} E\left[\beta^*\right]}{\Sigma_\delta}} = ||\Omega_{x_i,A} E\left[\beta^*\right]||_2 / \sqrt{\Sigma_\delta}, \tag{34}$$

where $\Sigma_\delta = \lim_{T\to\infty} TE\left[\hat{\delta}'\hat{\delta}\right]$ and $\hat{\delta} = \hat{\Psi} - \hat{\Phi}\beta$. Weighting the asymptotic bias of $\hat{\beta}$ by $\Omega_{x_i,A}$ (alternatively $\Omega_{z_i,A}$ if the counterfactual of interest is for a policy instrument) takes the quadratic norm of the bias in *units of the resulting counterfactual response* of the variable of interest. The denominator is the (scaled) asymptotic mean squared error in approximating the counterfactual rule across horizons (equivalently, the minimum distance objective function). As a result, the bias criterion measures the aggregate bias in the counterfactual response arising due to weak identification relative to the root mean square approximation error in implementing the counterfactual rule across horizons. This criterion is invariant to the scale of ζ_H. Essentially, it can be thought of as the (weighted) bias in the regression coefficients relative to the expected sum of squared residuals in a regression in IRF space. Since both the numerator and denominator depend on the unknown value, β, and on C, which cannot be consistently estimated, our test assesses the *worst case bias*, taken over all possible values for β and C. Appendix D develops expressions for this worst-case bias, adapting the results and proofs from Lewis and Mertens (2025a).

Following Lewis and Mertens (2025a), the null hypothesis that the bias exceeds some threshold can be tested using the statistic

$$g_{\min} = \text{mineval}\{\hat{\Delta}^{-\frac{1}{2}} m' \zeta_H \zeta_H' m \hat{\Delta}^{-\frac{1}{2}}/T\}, \tag{35}$$

$$= \text{mineval}\{\hat{\Delta}^{-\frac{1}{2}} \hat{\Phi}'\hat{\Phi} \hat{\Delta}^{-\frac{1}{2}}\}, \tag{36}$$

where $\hat{\Delta} = R_{K,H}'(\hat{\mathbf{W}}_2 \otimes I_H) R_{K,H}$, and $R_{K,H} = I_K \otimes \text{vec}(I_H)$. The second expression rewrites the test statistic in terms of the IRF coefficients, and is thus applicable to VARs as well. Appendix D characterises the limiting distribution of $g_{\min}$, which

is highly non-standard, and describes how critical values for that distribution can be constructed. These critical values depend only $\hat{\mathbf{W}}$. The test evaluates the null hypothesis that the asymptotic bias of $\hat{s}$, arising from weak identification, exceeds some tolerance threshold τ, conventionally set to $\tau = 0.10$ (10%), following Stock and Yogo (2005). If the null hypothesis is rejected, it indicates that weak identification is present, which, besides causing bias, also indicates robust inference methods should be used, since the distribution of $\hat{\beta}$ will be non-standard.

4 Inference

We now exploit the representation of $\hat{\beta}$ as a minimum distance estimator to develop inference procedures under both strong and weak identification, as well as tests for the implementatability of the counterfactual rule using the available instruments.

4.1 Inference for β

First, we impose a high-level assumption on the IRFs estimates:

Assumption 3.

$$\sqrt{T}\left(\begin{bmatrix}\text{vec}(\hat{\Pi}_{x,A})\\ \text{vec}(\hat{\Pi}_{z,A})\\ \text{vec}(\hat{\Omega}_{x,A})\\ \text{vec}(\hat{\Omega}_{z,A})\end{bmatrix} - \begin{bmatrix}\text{vec}(\Pi_{x,A})\\ \text{vec}(\Pi_{z,A})\\ \text{vec}(\Omega_{x,A})\\ \text{vec}(\Omega_{z,A})\end{bmatrix}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{M}). \tag{37}$$

and there exists an estimator $\hat{\mathbf{M}}$ *such that* $\hat{\mathbf{M}} \xrightarrow{p} \mathbf{M}$.

This assumption is consistent with standard properties of IRF estimators like VARs and LPs. Naturally, asymptotic normality then extends to $\hat{\Psi}$ and $\hat{\Phi}$:

$$\sqrt{T}\left(\begin{bmatrix}\text{vec}(\hat{\Psi})\\ \text{vec}(\hat{\Phi})\end{bmatrix} - \begin{bmatrix}\text{vec}(\Psi)\\ \text{vec}(\Phi)\end{bmatrix}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{W}), \tag{38}$$

where $\mathbf{W} = D_{\Psi\Phi}\mathbf{M}D'_{\Psi\Phi}$. $D_{\Psi\Phi}$ is a function of only the researcher's policy rule parameters,

$$D_{\Psi\Phi} = \begin{bmatrix}\tilde{A}_x^H & \tilde{A}_z^H & 0 & 0\\ 0 & 0 & I_K \otimes \tilde{A}_x^H & I_K \otimes \tilde{A}_z^H\end{bmatrix}. \tag{39}$$

The next assumption stipulates that identification is strong:

Assumption 4. $\Phi'\Phi$ *is fixed and full-rank.*

Under Assumptions 3 and 4, standard minimum distance results apply. Thus,

Proposition 2. *If Assumptions 3 and 4 hold and* $\beta \in \mathbb{B}$*, which is compact,*

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V_\beta), \tag{40}$$

where $V_\beta = D_\beta \mathbf{M} D_\beta'$ *and* D_β *is the Jacobian of* β *with respect to the IRFs of underlying variables; an explicit formula is given in Appendix C.*

Proposition 2 is a standard minimum distance result. Note that it holds regardless of whether Assumption 1 is maintained, but Assumption 1 is rather required for the estimates to be meaningful. The resulting distribution can be used to construct Wald confidence sets for β.

Without Assumption 4, Proposition 2 fails to hold. The limiting distribution of $\hat{\beta}$ is non-standard. Instead, inference can be conducted using weak identification robust methods, which are well-established for GMM and minimum distance problems (see, e.g., Stock and Wright (2000)). Before proceeding, we introduce the additional assumption that the counterfactual rule is *implementable*:

Assumption 5. *For the population weight vector,* β, $\Psi - \Phi\beta = 0$.

A counterfactual rule is implementatable if and only if the distance criterion can be set to exactly zero in population by β. The response of variables to the non-policy shock under the counterfactual rule is obtained as a linear combination of IRFs to the policy shocks. If there is no linear combination of those policy shock IRFs that matches the non-policy shock responses, the latter cannot be recovered.

Define the Anderson and Rubin (1949) statistic for β as

$$AR(b) = T f(b)' \hat{V}_f(b)^{-1} f(b), \tag{41}$$

where $\hat{V}_f(b) = \widehat{\text{var}(f(b))} = \left([1 \quad -b'] \otimes I_H\right) D_{\Psi\Phi} \hat{\mathbf{M}} D_{\Psi\Phi}' \left([1 \quad -b']' \otimes I_H\right)$. Then

Proposition 3. *Under Assumptions 3, and 5, provided* $V_{AR}(\beta)$ *is positive definite,*

$$AR(\beta) \xrightarrow{d} \chi^2_H. \tag{42}$$

As a result of Proposition 3, a confidence set for β can be constructed via test inversion. In other words, the confidence set is defined as

$$CS^{AR}_\beta(\alpha) = \{b : AR(b) \leq \chi^2_H(\alpha)\}. \tag{43}$$

Such a set can be constructed via test inversion, where the set of values b for which $AR(b) \leq \chi^2_H(\alpha)$ is collected, often over a grid.

In practice, H may be considerably larger than K (in McKay and Wolf (2023), $H = 101$ and $K = 2$). In such highly over-identified models, it is well-documented that the power properties of the Anderson and Rubin (1949) statistic may be poor. We thus also consider Kleibergen's (2005) KLM statistic, which addresses these concerns:

Proposition 4. *Define*

$$KLM(\beta) = T^{-1} \tilde{q}(\beta) \left[\hat{\mathcal{D}}(\beta)' \hat{V}_f(\beta)^{-1} \hat{\mathcal{D}}(\beta)\right]^{-1} \tilde{q}(\beta)', \tag{44}$$

where $\tilde{q}(\beta) = f(\beta)'\hat{V}_f(\beta)^{-1}\hat{\mathcal{D}}(\beta)$, *and* $\hat{\mathcal{D}}(\beta)$ *is the derivative of the AR statistic with respect to* β. *Then under Assumptions 3-5,*

$$KLM(\beta) \xrightarrow{d} \chi^2_K. \tag{45}$$

Explicit expressions for $\hat{\mathcal{D}}(\beta)$ are given in Appendix C and specialise Theorem 1 of Kleibergen (2005) to the present setting. $CS^{KLM}_{\beta}(\alpha)$ is defined analogously to $CS^{AR}_{\beta}(\alpha)$.

Remark 1. The requirement that $V_f(\beta)$ be positive definite in Proposition 3 is non-trivial. For a VAR, $\mathbf{M}$ can only be positive definite if $H = |\mathcal{H}| \leq pn^2 + (K+1)n$. Since the asymptotic normal distribution is a second-order approximation, the rank of $\mathbf{M}$ and thus that of $V_f(\beta) = D_{\Psi\Phi}\mathbf{M}D'_{\Psi\Phi}$ cannot exceed that of the variance matrix of the estimated VAR parameters used to compuate the IRFs (autoregressive parameters and identified columns of contemporaneous response matrix). If the asymptotic variance of $f(\beta)$ is not positive definite, it is not possible to construct the Anderson and Rubin (1949) statistic, or the KLM statistic (which requires the same inversion). This fact will play an important role in the selection of $\mathcal{H}$. In practice, the maximum number of horizons for which $V_f(\beta)$ is full rank, $\bar{H}$, can be considerably lower still, and depend on the structure of the counterfactual considered. This is the case in our empirical illustration, where, although the IRFs depend on up to 217 parameters, the maximum number of horizons for which $V_f(\beta)$ is full rank ranges from 34 to 42 across counterfactuals.

Remark 2. Even if Wald inference is to be conducted, for which $\mathbf{M}$ in principle need not be invertible, the above remark illustrates that it is ill-advised to use $|\mathcal{H}| > pn^2 + (K+1)n$ horizons to estimate β. To a second-order approximation, the linearly independent information contained in horizons $\mathcal{H}$ is of rank at most $pn^2+(K+1)n$, since the IRFs are constructed as a function of $pn^2+(K+1)n$ VAR parameters. Therefore, using additional horizons can only add linearly dependent information, akin to adding irrelevant instruments.

4.2 Testing for implementability of a counterfactual rule

We argued above that a counterfactual rule is implementable on the basis of the K instruments in m if and only if Assumption 5 holds. An immediate consequence is that if $H > K$, testing the over-identifying restrictions used to estimate β is a test of implementability. In the case of strong identification, this amounts to Hansen's $J-$test. Under weak identification, the equivalent procedure is to determine whether the Anderson and Rubin (1949) confidence set is empty. Both procedures use the same test statistic, but different critical values:

Proposition 5. *To test the null hypothesis of implementability, while maintaining Assumption 1, compute the test statistic* $\hat{J} = AR(\hat{\beta})$. *If Assumptions 3-5 hold*

(strong identification), then $\hat{J} \xrightarrow{d} \chi^2_{H-K}$. *If Assumption 4 does not hold (weak identification), then* $\hat{J} \xrightarrow{d} \chi^2_H$.

While this result describes a standard test of over-identifying restrictions, the insight that rejection of the null hypothesis implies the counterfactual rule is not implemented is novel; this condition was not otherwise formally testable.

4.3 Inference for counterfactuals

The ultimate object of interest is the vector $\Xi \equiv \Pi_{i,\tilde{A}}$, the IRF of variable i to the non-policy shock under counterfactual rule $\tilde{A}$. We now extend the above results to provide inference methods for Ξ.

To proceed horizon-by-horizon, we augment the distance function $f(\cdot)$ based on the additional equation

$$\Xi^h - \Pi^h_{i,A} - \Omega^h_{i,A}\beta = 0, \tag{46}$$

to yield the new function

$$g(b,s) = \begin{pmatrix} \hat{\Psi} + \hat{\Phi}b \\ s - \hat{\Pi}^h_{i,A} - \hat{\Omega}^h_{i,A}b \end{pmatrix}. \tag{47}$$

The resulting distribution follows.

Proposition 6. *Under Assumptions 3 and 4,*

$$\sqrt{T}(\text{vec}(\hat{\Xi}^h) - \text{vec}(\Xi^h)) \xrightarrow{d} \mathcal{N}(0, V_\Xi), \tag{48}$$

where $V_\Xi = D_\Xi \mathbf{M} D_\Xi'$.

D_Ξ is the Jacobian of Ξ^h with respect to the IRFs of underlying variables, and explicit expressions are given in Appendix C.

Under weak identification, we define the AR statistic analogously to before,

$$AR(\beta, \Xi^h) = g(\beta, \Xi^h)' \hat{V}_g(\beta, \Xi^h)^{-1} g(\beta, \Xi^h) \xrightarrow{d} \chi^2_{H+1}, \tag{49}$$

where explicit expressions for $\hat{V}_g(\beta, \Xi^h)$ are provided in Appendix C. Similarly, the KLM statistic is now

$$KLM(\beta, \Xi^h) = T^{-1}\tilde{q}(\beta, \Xi) \left[\hat{\mathcal{D}}(\beta, \Xi^h)' \hat{V}_g(\beta, \Xi^h)^{-1} \hat{\mathcal{D}}(\beta, \Xi^h)\right]^{-1} \tilde{q}(\beta, \Xi^h)' \xrightarrow{d} \chi^2_{K+1}, \tag{50}$$

where $\tilde{q}(\beta) = g(\beta, \Xi^h)' \hat{V}_g(\beta, \Xi^h)^{-1} \hat{\mathcal{D}}(\beta, \Xi^h)$ and explicit expressions for $\hat{\mathcal{D}}(\beta, \Xi^h)$ are provided in Appendix C. Confidence sets $CS^{AR}_\Xi(\alpha)$ and $CS^{KLM}_\Xi(\alpha)$ can be constructed analogously to those for β.

Remark 3. The augmented distance function introduced above highlights limitations to what types of counterfactuals can be considered meaningful targets for

inference in the McKay and Wolf (2023) framework. Counterfactual rules that depend on a single variable – like the output gap targeting and interest rate peg counterfactuals in Section 5 – present a conceptual challenge that becomes stark when asymptotic inference is considered. In the former, $\Psi = \Pi_{x_i,A}$ and $\Phi = \Omega_{x_i,A}$, where x_i is the output gap, and in the latter, $\Psi = \Pi_{z_i,A}$ and $\Phi = \Omega_{z_i,A}$, where z_i is the nominal interest rate. Considering the first case, the augmented distance function includes the following two conditions for any horizon $h \in \mathcal{H}$:

$$\Pi^h_{x_i,A} + \Omega^h_{x_i,A}\beta = 0 \tag{51}$$

$$\Xi^h - \Pi^h_{x_i,A} - \Omega^h_{x_i,A}\beta = 0. \tag{52}$$

Clearly, these two conditions cannot both hold unless $\Xi^h = 0$! Therefore, if the researcher believes that the counterfactual rule is implementable, then she immediately knows *by assumption* the true value of the counterfactual response of x_i: zero! Any effort to report inference for the counterfactual path is disingenuous, since any values besides zero contained in a confidence interval are known to be due to sampling error in finite samples or non-implementablility of the rule asymptotically. Much like the test for implementability itself, this fact becomes clear in a frequentist framework.

Remark 4. Even if a researcher still wanted to report confidence intervals for the counterfactual responses of variables appearing in single-variable counterfactual rules, it is not possible to do so at horizons used for estimation. The variance matrix, V_g, is rank deficient. This means that both the AR and KLM statistics are undefined.

Remark 5. Several comments are in order regarding inference for optimal policy counterfactuals. First, interpretation of the robust confidence sets may require caution. This is because $\tilde{A}^H_x$ essentially re-scales and sums the endogenous regressors, but the test statistics (and their critical values) are invariant to invertible re-scalings and linear combinations. Thus, the confidence sets may not, in general, contain the point estimate for the optimal policy path (this is the case for some horiozns in Figure 8 below). If the researcher is nevertheless interested in conducting inference under a particular weighting, non-standard critical values can be computed.[3] Second, certain objective functions may lead to certain counterfactual responses not being feasible targets for asymptotic inference, as in Remark 3. For instance, the objective studied by McKay and Wolf (2023), (26), results in a rank-deficient asymptotic covariance matrix when Ξ^h is the response of the output gap at some horizon h, since this variable enters the loss function directly. There is a more subtle issue for inflation responses, since the value entering the objective for each

[3]For instance, AR confidence sets respecting a particular weighting can be constructed by replacing $\hat{V}_g(\beta, \Xi^h)$ in (49) by the identity and simulating critical values by drawing N $(H+1)\times 1$ random normal vectors from the distribution $\kappa^{(\mathfrak{n})} \sim \mathcal{N}(0, V_g(\beta, \Xi^h))$ and using the α-quantile of $\kappa^{(\mathfrak{n})\prime}\kappa^{(\mathfrak{n})}$, $\mathfrak{n} = 1, \ldots, N$ as a critical value.

period considers the average level of inflation over several quarters; the eigenvalues of the asymptotic covariance matrix may become small depending on the window over which inflation is averaged.

Remark 6. It is straightforward to extend the results above to conduct inference for multiple elements or indeed all of Ξ (and even the counterfactual responses of different variables) by augmenting the distance function with additional equations. We opt to focus on pointwise inference, horizon by horizon, since robust inference is likely to be the norm, and adding further equations to the distance function risks further proliferation of nuisance paramters and thus deterioration of testing power using projection methods if not all parameters are simultaneously the subject of a joint hypothesis.

Remark 7. To construct a pointwise identification-robust confidence interval for Ξ^h, a test inversion approach can be applied to $AR(\beta, \Xi^h)$ or $KLM(\beta, \Xi^h)$. This method consists of finding the values of (β, Ξ^h) (on a grid, say) such that the test statistic takes a value less than or equal to the relevant critical value. Collecting the values that cannot be rejected in this manner yields a joint confidence set for (β, Ξ^h). Constructing a confidence *interval* for Ξ^h proceeds using the *projection method*, which obtains the range of values for Ξ^h for which there exists *some value for* β such that the test statistic does not exceed the critical value. Formally, for the case of the $AR(\beta, \Xi^h)$ statistic, the projection interval for Ξ^h is the set

$$CI^{AR}_{\Xi^h}(\alpha) = \{s : \min_{b \in \mathbb{B}} AR(b, s) \leq \chi^2_{H+1}(\alpha)\}, \tag{53}$$

which can either be constructed on a grid or via numerical minimisation.

4.4 Selection of horizons for estimation

The previous sections highlight the central role that the selection of horizons plays in estimating macroeconomic counterfactuals. First, identification is likely to be weak, with the choice of horizons essentially standing in for the choice of instruments. As a result, using horizons with near-zero, collinear, or imprecisely estimated IRFs is unadvisable. Second, while in theory all feasible horizons should be used to ensure the counterfactual rule is implemented in all time periods, the approach of McKay and Wolf (2023), using all horizons out to some very large number, is not possible if frequentist inference is to be conducted based on a VAR, since the asymptotic distribution of the IRFs will be degenerate. Even for estimation, this result indicates that – to a second-order approximation – using more horizons than $H^{\max}$, the number of VAR parameters, is tantamount to adding irrelevant instruments (Remark 2). This second fact, that all horizons *cannot* be used, implies that some method for horizon selection is not just desirable, but indeed necessary, for a frequentist implementation. It likewise implies that the choice of horizons should be considered carefully in Bayesian implementations.

We propose to select the set of horizons for estimation using an information criterion-type approach that is based on an econometrician's loss function, trading off bias and variance. Recall that the question of which horizons to use to estimate β is a separate question from the horizons for which the researcher wishes to construct counterfactual responses. A horizon h that is not part of $\Psi(\mathcal{H})$ and $\Phi(\mathcal{H})$ can easily still be used to compute counterfactuals following (15).

We measure bias based on the difference between the estimated counterfactual responses computed using a candidate set of horizons, $\mathcal{H}$, and the estimated responses using horizons $0, \ldots, \bar{H}-1$, the maximum number feasible in a given VAR (the maximum such that $V_f(\beta)$ is full rank). We evaluate the difference over $0, \ldots, \bar{H}-1$. For a LP, there is not so obvious a choice for $\bar{H}$, but it seems natural to set it to the value that would apply to the equivalent VAR (based on results in Plagborg-Møller and Wolf (2021)). Given the structure of the counterfactual estimator, for a variable i, this difference in counterfactual paths is given by

$$\hat{\Omega}_{i,A}\left(\hat{\beta}(\mathcal{H}) - \hat{\beta}(\bar{H})\right), \tag{54}$$

where $\hat{\beta}(\mathcal{H})$ denotes the estimate based on the horizons in $\mathcal{H}$ and $\hat{\beta}(\bar{H})$ denotes that based on $0 \ldots, \bar{H}-1$. We consider the average ℓ_2 norm of this bias over the d variables whose counterfactuals are of interest.

Typically, (some function of) the variance of an estimator, $\hat{\beta}$, would be used to compute an information criterion. However, we want to tailor our approach to the most empirically probable setting – weak identification – where the distribution of $\hat{\beta}$ cannot be characterised. Thus, to compute an identification-aware bias-variance tradeoff, we propose to use the volume of the AR or KLM confidence set for β.

The resulting loss function, based on confidence set of type $c \in \{AR, KLM\}$, is

$$\mathcal{L}(\mathcal{H}) = w \times \frac{1}{d}\sum_{i=1}^{d} \|\hat{\Omega}_{i,A}\left(\hat{\beta}(\mathcal{H}) - \hat{\beta}(\bar{H})\right)\|^2 + (1-w) \times |CS^c_\beta(\alpha, \mathcal{H})|, \tag{55}$$

where $|CS^c_\beta(\alpha, \mathcal{H})|$ denotes the volume of $CS^c_\beta(\alpha, \mathcal{H})$, where we now make the dependence of the confidence set on $\mathcal{H}$ explicit. In the common case of $K = 2$ (as in our empirical study), this is simply the area of the level-α robust confidence set. In practice, we calculate this by considering a grid of angles from 0 to 2π, computing the radius of the confidence set along each direction from the point estimate, and integrating to approximate the area.

While $|CS^c_\beta(\alpha, \mathcal{H})|$ could be measured for the counterfactual response of interest at a particular horizon, Ξ^h, (rather than the intermediate estimand β), we prefer to use β so that $\mathcal{H}$ does not vary over the different responses or indeed horizons reported for a single counterfactual rule. Averaging the volume over all horizons and variables of interest d is computationally infeasible (as discussed below, the proposed apporach is already computationally intensive). Moreover, in our experi-

mentation, we found that the horizons selected based on β and those selected using the responses at various horizons and variables were remarkably similar. Thus, it makes sense to study β and hold the estimation specification constant for all results reported for a single counterfactual rule.

We select the horizons for estimation as the set of horizons, $\mathcal{H}^*$, that minimises the loss (55) amongst a set of possible horizon combinations $\mathbb{H}$:

$$\mathcal{H}^* = \underset{\mathcal{H}\in\mathbb{H}}{\operatorname{argmin}}\ \mathcal{L}(\mathcal{H}). \tag{56}$$

Defining $\mathbb{H}$, the set of horizons considered, requires care. As discussed by Hall et al. (2012) for their proposed information criterion to select horizons for IRF matching exercises, the computational burden of considering all possible combinations of horizons quickly becomes infeasible. To this end, we propose to consider all combinations of $K, K+1, \ldots, |\mathcal{H}|^{\max}$ horizons from amongst the set $\{h = 0, \ldots, \bar{h}\}$, where $|\mathcal{H}|^{\max} \leq \bar{H}$ and $\bar{h} \leq \bar{H} - 1$. In our empirical application, for inference based on the KLM statistic (our baseline) we set $\bar{h} = 19$ and increase $|\mathcal{H}|^{\max}$ until we obtain an interior solution (i.e. strictly fewer than $|\mathcal{H}|^{\max}$ horizons are selected). These choices are made partially for computational reasons; already, this requires searching over up to $\prod_{j=2}^{13} {}_{20}C_j = 988,095$ combinations of horizons.[4] However, experimentation suggests that further extending the number of horizons or the range of horizons considered does not meaningfully change the horizons selected. Note that while Hall et al. (2012) reduce the dimension of their problem by considering only consecutive sequences of horizons starting from zero, such an approach would be ill-advised in the present setting, since consecutive horizons are the most likely to contain redundant information, and, depending on the policy shocks considered, the first horizon(s) may exhibit near-zero responses due to response lags. We set $w = 0.75$ to prioritise small bias and approximating the counterfactual well across a wide range of horizons over more precise inference (stronger identification).

Note that although the loss function is (indirectly) based on a test statistic, this selection rule does not introduce something resembling a pre-testing problem. This is because it is used to select between a set of pre-defined models and their associated estimates (including confidence sets), rather than being part of a sequential testing procedure in which the final estimation and inference step depends on the outcome of an earlier test, conducted on the same sample. All of the $\dim(\mathbb{H})$ preliminary confidence sets for β are valid level-α sets (associated with similarly valid intervals for each $\Pi^h_{i,\tilde{A}}$), and our criterion simply selects amongst the associated models. Moreover, the volume of the robust confidence set is not actually a testing outcome itself, and is not directly informative for the outcome of any hypothesis test for the counterfactual responses (or even β) since the volume contains no information on the *location* of the set's volume relative to any value of interest for β, and certainly

[4]For the AR statistic, for which the computational burden is lower, we set $H^{max} = 6$ and $\bar{h} = 23$, resulting in 190,026 combinations.

not $\Pi^h_{i,\tilde{A}}$. As a result, this formal procedure does not facilitate specification search for significant results.

4.5 Simulation evidence

To evaluate the performance of our estimation and inference procedures in practice, we conduct simulations using data generated from a textbook three-equation NK model, based on Galí (2015). We calibrate the parameters as described therein, except for setting the coefficient of relative risk aversion to 2 (to accommodate risk aversion) and setting the volatilties of cost-push and demand shocks to 1 and that of the monetary policy news shocks to 0.5 (so that the monetary policy shock explains relatively less variation, consistent with weak identification). We consider models with two sets of dynamics. In the first, we model the cost-push shock as an MA(4), and in the second as an MA(1); in the first case, this means $K = 5$ policy shocks are required to implement the counterfactual rule, while in the latter, two policy news shocks suffice to implement it. The p^{th} MA coefficient is 0.95^p. All policy news shocks enter additively (with unit coefficients) in the Taylor Rule, as in (4). We assume that the instruments available to the econometrician are generated as

$$e_t = \varsigma \epsilon_t + \sqrt{1-\varsigma^2} w^e_t \tag{57}$$

$$m_{it} = \varsigma \nu_{it} + \sqrt{1-\varsigma^2} w^m_{it}, \; i = 1, \ldots, 5, \tag{58}$$

where $\varsigma = 0.6$ and measurement errors $(w^e_t, \; w^{m\prime}_t)' \sim \mathcal{N}(0, I_6)$, following the design in Lewis and Mertens (2026). We simulate 5,000 samples for $T = 150$ (approximating the sample length in our quarterly empirical application) and $T = 500$ (approximating 40 years of monthly data). We consider the response of output under a counterfactual Taylor Rule, where instead of responding to inflation with $\phi = 1.5$, the policymaker responds more hawkishly, with $\tilde{\phi} = 3.0$.

Table 1 reports the results. The top panel considers the MA(4) cost-push shock model, where the econometrician has five monetary policy news shocks available, $K = 5$. The middle and bottom panels consider the MA(1) model where the econometrician has access to only two shocks, $K = 2$, but these are adequate to fully implement the counterfactual. The first two panels use $H = 5$, while the last panel selects horizons to minimise (55). The left panels report results for $T = 150$ and the right for $T = 500$. For each panel, the top row reports the bias for β, averaged across horizons, and by horizon. The second row reports the RMSE for β, averaged across horizons, and by horizon. The next rows report Wald, AR, and KLM coverage for nominal 95% intervals. Under "Test statistics", we report the average value of g_{min} and critical values for weak identification and the average value of $\hat{J}$ for implementability ($cv = 11.07$, or 7.81 assuming strong identification).

There are few clear patterns to bias; when $K = 5$, bias does not decrease with

larger samples, but with $K = 2$ it does.[5] Bias also lower for $K = 2$ than for $K = 5$, although the IRFs of the two models are different, so this is not a direct comparison. The RMSE falls with sample size throughout, both jointly and for all horizons except for one in the $K = 5$ model.

The Wald tests offer poor coverage for both models; the nominal level is 95%. Coverage for β is only 0.61 ($K = 5$) and 0.24 ($K = 2$), while joint coverage for the counterfactual response of output at all horizons is 0.81 and 0.74, respectively. Coverages can be considerably lower still at some horizons – especially short ones – while they tend towards 1 at longer horizons. In either case, inference based on standard asymptotics is poorly sized. When the model is just-identified ($K = H = 5$), the AR and KLM statistics coincide; coverage for both β and the counterfactuals is very close to the nominal level of 0.95. When $K = 2$, the statistics differ, but both over coverage equal or close to the nominal level. This is true whether H is fixed at 5 or chosen optimally.

We report the average value of $g_{\min}$ across samples and the associated critical value across samples. Doing so makes clear that the $K = 5$ model is extremely weakly identified, with a test statistic of essentially zero, and critical values on the order of 8 ($T = 150$) and 6 ($T = 500$). For $K = 2$, the test statistics rise to around 0.5-0.7 on average, with critical values over 3, still representing quite weak identification. In the bottom panels, we can also test for implementability, since $H > K$. We report empirical rejections rates $(\hat{\alpha}_s(J), \hat{\alpha}_w(J)$ using critical values that valid under both strong and weak identification; for both sets of values, the test is conservative, rejecting below the nominal 5% level, as has previously been documented in the case of weak identification (e.g., Andrews et al. (2019)). We also report the average value of $|\mathcal{H}^*_{KLM}|$. These remain very close to 5, because in the majority of samples, there is at least one direction in which the KLM set if effectively unbounded for all combinations of horizons; as a result, the criterion simply selects the set of horizons that minimises the bias relative to $\bar{H}$, which is $\bar{H} = 5$ itself.

In Appendix B, we report simulation results for an alternative calibration with stronger identification; these demonstrate performance when there is more variation in $\mathcal{H}^*_{KLM}$, as well as how our other procedures vary with identification strength. They show that as identification strengthens, non-robust inference procedures remain poorly approximated by standard distributions, our horizon-selection criteria successfully chooses horizons that yield stronger identification, our bias-based first-stage test has non-trivial power in detecting strong identification, and the empirical size of our test for implementability approaches its nominal level.

[5] In additional experiements, bias does fall as T is taken considerably larger.

Table 1: Performance in simulations

$H=5, K=5$					$T=150$							$T=500$		
	β	Joint	$h=0$	1	2	3	4	β	Joint	$h=0$	1	2	3	4
Bias	2.64	5.29	11.60	4.23	-1.39	7.14	4.86	1.83	6.31	3.94	9.81	8.01	6.89	2.89
RMSE	124.19	364.04	462.77	128.48	410.95	445.19	372.82	98.72	224.95	182.54	241.69	224.68	217.96	257.90
Wald coverage	0.61	0.81	0.87	0.91	0.95	0.99	1.00	0.85	0.90	0.93	0.95	0.97	0.99	1.00
AR/KLM coverage	0.95	0.94	0.93	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95
Test statistics			$g_{min}^{avg}=0.01$		$cv=8.57$					$g_{min}^{avg}=0.02$		$cv=6.40$		
$H=5, K=2$					$T=150$							$T=500$		
	β	Joint	$h=0$	1	2	3	4	β	Joint	$h=0$	1	2	3	4
Bias	1.31	0.40	1.30	0.67	0.04	-0.01	-0.01	0.90	0.26	0.84	0.43	0.03	0.01	0.01
RMSE	1.45	0.99	1.57	1.11	0.78	0.75	0.76	1.16	0.79	1.14	0.93	0.63	0.61	0.63
Wald coverage	0.24	0.74	0.66	0.93	1.00	1.00	1.00	0.52	0.76	0.73	0.94	0.99	1.00	0.99
AR coverage	0.95	0.95	0.94	0.95	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95
KLM coverage	0.94	0.94	0.92	0.94	0.94	0.94	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.95
Test statistics	$g_{min}^{avg}=0.47$		$cv=3.61$		$\hat{\alpha}_s(J)=0.00$		$\hat{\alpha}_w(J)=0.00$	$g_{min}^{avg}=0.69$		$cv=3.14$		$\hat{\alpha}_s(J)=0.01$		$\hat{\alpha}_w(J)=0.00$
$H=\lvert\mathcal{H}_{KLM}^*\rvert, K=2$					$T=150, \lvert\mathcal{H}_{KLM}^*\rvert=4.98$							$T=500, \lvert\mathcal{H}_{KLM}^*\rvert=4.88$		
	β	Joint	$h=0$	1	2	3	4	β	Joint	$h=0$	1	2	3	4
Bias	1.32	0.40	1.31	0.68	0.03	-0.01	0.00	0.88	0.26	0.83	0.42	0.04	-0.01	-0.01
RMSE	1.45	0.99	1.55	1.09	0.78	0.76	0.75	1.26	0.87	1.25	1.06	0.67	0.65	0.73
AR coverage	0.96	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.95	0.95	0.96	0.95	0.95	0.95
KLM coverage	0.96	0.94	0.95	0.96	0.96	0.96	0.96	0.96	0.94	0.95	0.96	0.96	0.96	0.96
Test statistics	$g_{min}^{avg}=0.47$		$cv=3.63$		$\hat{\alpha}_s(J)=0.00$		$\hat{\alpha}_w(J)=0.00$	$g_{min}^{avg}=0.73$		$cv=3.20$		$\hat{\alpha}_s(J)=0.01$		$\hat{\alpha}_w(J)=0.00$

Notes: The table reports simulation performance for the three-equation NK model described in the text, for a Taylor Rule counterfactual output response. The top panel considers a model with $K=5$ monetary policy shocks, all of which are available to the econometrician, while in the middle and bottom, only two shocks are available, but these suffice to implement the counterfactual rule. For the top two panels, five horizons are used for estimation, and for the bottom, horizons are chosen to minimise (55). The left panels consider $T=150$, and the right $T=500$. For each panel, the top row reports the bias for β, averaged across horizons, and by horizon. The second row reports the RMSE for β, averaged across horizons, and by horizon. The next rows report Wald, AR, and KLM coverage for nominal 95% intervals. Under "Test statistics", we report the average value of g_{min} and critical values for weak identification as well as the empirical rejection rates of our test for implementability using critical values valid under both strong and weak identification ($\hat{\alpha}_s(J), \hat{\alpha}_w(J)$, respectively).

5 Empirical illustration

We now apply our tests, inference procedures, and the lessons for empirical design from our framework to the four counterfactual rules and optimal policy problem considered by McKay and Wolf (2023). To be specific, these consist of:

1. Output Gap (OG), $y_t = 0$
2. Taylor Rule (TR), $i_t = 0.5i_{t-1} + 0.5(1.5\pi_t + y_t)$
3. Interest rate peg (PEG), $i_t = 0$
4. Nominal GDP target (NGDP), $\pi_t + (y_t - y_{t-1}) = 0$
5. Optimal Average inflation targeting (OPT), $\mathfrak{L} = \bar{\pi}' W \bar{\pi} + y' W y$, $W = \text{diag}(1, \beta, \beta^2, \ldots)$, $\beta = 1/1.01$.

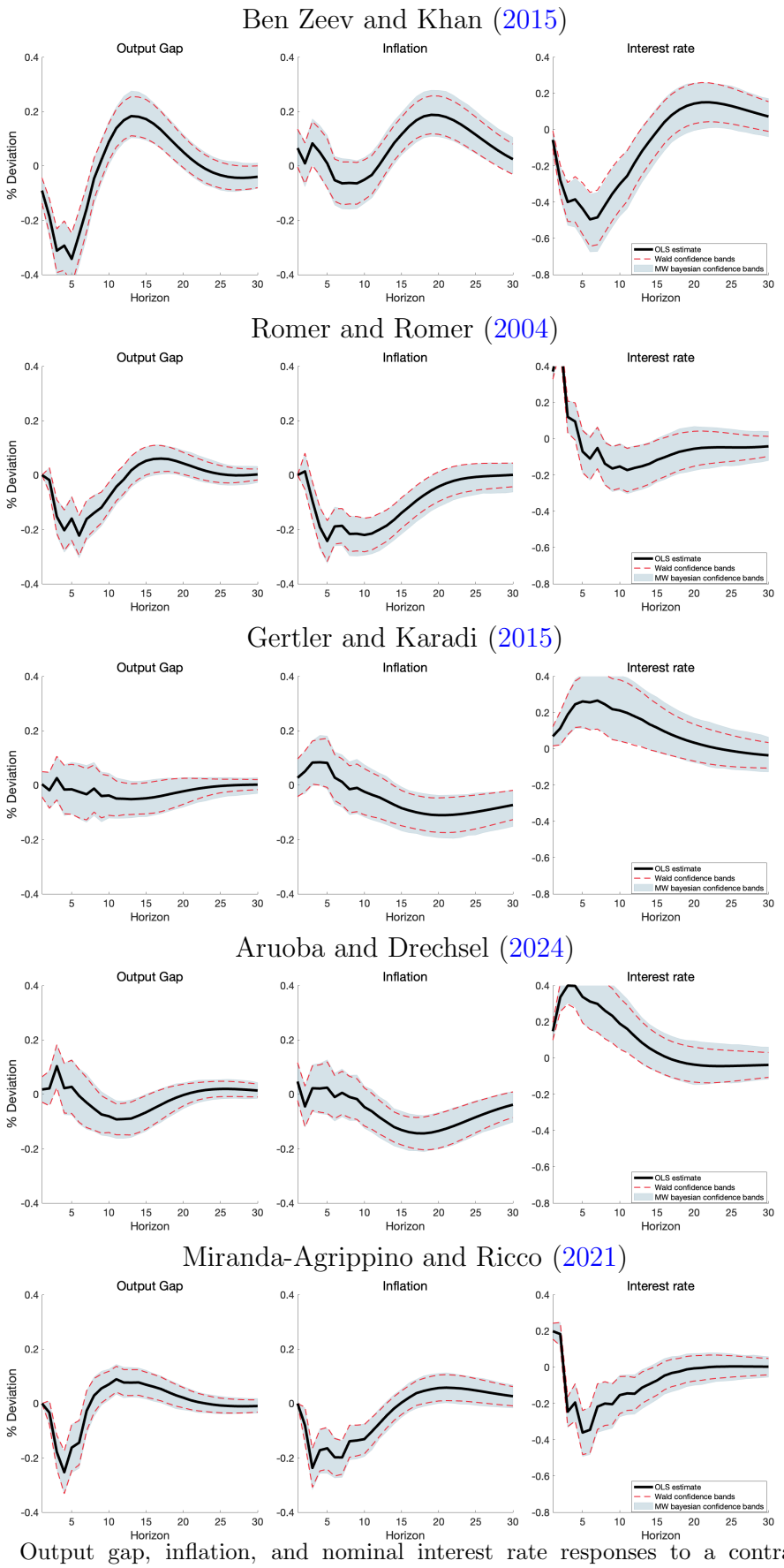
In each case, the non-policy shock of interest is an investment-specific technology shock, $x_t = (y_t, \pi_t)'$, where y_t is the output gap and π_t is inflation, and $z_t = i_t$, the nominal interest rate. The analysis is based on two instruments for monetary policy shocks, so $K = 2$. The technology shock instrument is that of Ben Zeev and Khan (2015). For monetary policy, we follow McKay and Wolf (2023) and consider two pairs of shocks: Romer and Romer (2004) and Gertler and Karadi (2015), or Aruoba and Drechsel (2024) and Miranda-Agrippino and Ricco (2021).

McKay and Wolf (2023) use two different VARs to estimate the IRFs to the technology shock and the monetary policy shocks. For frequentist inference, the covariance of the IRF estimates is a crucial ingredient. As a result, we instead estimate all IRFs jointly using a single VAR. We estimate the IRFs using the internal instruments approach. We order the variables as the technology shock, the first monetary policy shock, output gap, inflation, commodity prices, the second monetary policy shock, and the Fed Funds rate. Ordering one policy shock before the aggregate variables and one after follows McKay and Wolf (2023). Depending on the shock pair, the first monetary policy shock is either Gertler and Karadi (2015) or Aruoba and Drechsel (2024), and the second shock is either Romer and Romer (2004) or Miranda-Agrippino and Ricco (2021). As in McKay and Wolf (2023), the VAR contains a linear time trend and 4 lags of all variables. The sample spans 1969:Q1 to 2007:Q4.

Figure 3 plots the OLS estimates of the IRFs to each shock, with a 95% confidence interval, along with their Bayesian counterparts (posterior median), applying McKay and Wolf's (2023) code to our unified VAR. For both point estimation and inference, the choice of estimation framework clearly makes little difference for the underlying IRFs. The baseline analysis in McKay and Wolf (2023) uses the Romer and Romer (2004) and Gertler and Karadi (2015) shocks, with results using Aruoba and Drechsel (2024) and Miranda-Agrippino and Ricco (2021) reported for robustness. However, careful examination of the IRFs in Figure 3 suggests that, while the IRFs generally have similar profiles, as noted by McKay and Wolf (2023), the responses to the Gertler and Karadi (2015) shock are somewhat less pronounced than those to the Aruoba and Drechsel (2024) shock, especially for the output gap. In addition, the estimates appear to be less precisely. These properties indicate that the former is likely to offer relatively less sharp identification. To formally confirm this visual intuition, we conduct first stage tests for each rule considered based on the full possible range of horizons, $\mathcal{H} = 0, \ldots, \bar{H}$. Table 2 reports the results. It confirms that while identification is weak regardless of the instruments used, with the exception of PEG, the test statistics are larger relative to the respective critical values for the latter shock combination for all variables of interest. We thus adopt Aruoba and Drechsel (2024) and Miranda-Agrippino and Ricco (2021) as our baseline for the remainder of our analysis, since these shocks appear to offer somewhat sharper identification. We report results using the Romer and Romer (2004) and Gertler and Karadi (2015) shocks instead in Appendix A.4.

We next turn to the selection of horizons used to approximate each counterfactual. Figure 3 reports horizons out to 30 quarters. At this point, all responses have converged to zero. However, recall that McKay and Wolf (2023) use $\mathcal{H} = \{0, 1, \ldots, 99\}$ for their analysis. Not only is this not feasible in the frequentist framework (see Remark 1), but it is not desirable either. Each additional horizon beyond approximately 30 is like adding a completely irrelevant instrument, which

Figure 3: IRFs to non-policy and policy shocks



Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock (row 1), the Romer and Romer (2004) monetary policy shock (row 2), the Gertler and Karadi (2015) monetary policy shock (row 3), the Aruoba and Drechsel (2024) monetary policy shock (row 4), and the Miranda-Agrippino and Ricco (2021) monetary policy shock (row 5). The solid black line reports the point estimate, with the 68% Wald confidence set shaded in red. The blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Table 2: Comparison of shock measures and identification strength

	Rules	$\bar{H}$	Test statistic	Critical value		
				y	π	i
RR and GK shocks	OG	41	0.24	6.28	7.65	8.61
	TR	43	1.77	6.10	7.36	8.53
	PEG	41	1.07	4.99	5.82	6.54
	NGDP	43	1.19	5.40	6.51	7.44
	OPT	34	0.72	4.99	6.02	6.81
AD and MAR shocks	OG	39	0.34	6.37	6.94	9.59
	TR	42	2.46	6.04	6.62	9.10
	PEG	39	0.86	5.10	5.49	7.09
	NGDP	41	1.57	5.31	5.80	7.95
	OPT	34	1.13	5.23	5.68	7.75

Notes: Results of the first-stage test for weak identification using the horizons $0, \ldots, \bar{H} - 1$ for each counterfactual rule. The first panel reports the results using the Romer and Romer (2004) and Gertler and Karadi (2015) shocks and the second using the Aruoba and Drechsel (2024) and Miranda-Agrippino and Ricco (2021) shocks. The first column of each panel reports the test statistic and the second column reports the 5% critical value for a bias tolerance of 10%.

only serves to add noise to the "first-stage" due to over-fitting. However, the results in Table 2 suggest further that using all horizons from 0 up to the feasible limit, $\bar{H}$, provides very weak identification. Figures 9-13 in Appendix A.1 report the associated counterfactuals for reference. As expected, the robust confidence sets are extremely wide. For all but one of the counterfactual responses, the KLM set is *unbounded*, at least out to magnitudes of reasonable interest. For no counterfactual rule is it possible to reject that any variable responds differently under the imposed rule than it does under the rule prevailing in the data. That is – the exercise is entirely economically uninformative if robust inference is conducted with $\mathcal{H} = \{0, 1, \ldots, \bar{H} - 1\}$.

On the basis of these results, it might appear that – once the presence of weak identification is acknowledged and accounted for – nothing can be learned from from the counterfactual exercises in McKay and Wolf (2023). However, applying our procedure to select horizons for estimation can significantly strengthen identification. The top panel of Table 3 compares the results of the first-stage test using $\mathcal{H} = \{0, 1, \ldots, \bar{H} - 1\}$ (as previously reported in Table 2) with those based on the optimal set of horizons following our selection rule, reported in the next two panels (based on AR and KLM sets, respectively). Selecting horizons based on the AR sets tends to select a small number of horizons, ranging from two for TR and OPT to five for PEG and NGDP, and results in larger test statistics for three of the five rules considered, although the critical values tend to be somewhat larger too. Selecting horizons based on the KLM sets selects strictly more horizons, ranging from four for OPT to 12 for TR, resulting in smaller test statistics than using $\mathcal{H}^*_{AR}$,

Table 3: Diagnostic results and horizon choice

	Rules	H Length	Test statistic	Critical value		
				y	π	i
First-stage test using $\mathcal{H} = \{0, 1, \ldots, \bar{H} - 1\}$	OG	39	0.34	6.37	6.94	9.59
	TR	42	2.46	6.04	6.62	9.10
	PEG	39	0.86	5.10	5.49	7.09
	NGDP	41	1.57	5.31	5.80	7.95
	OPT	34	1.13	5.23	5.68	7.75
First-stage test using $\mathcal{H}^*_{AR}$	OG	4	0.41	9.50	10.46	15.12
	TR	2	3.28	17.70	19.85	31.56
	PEG	5	4.27	7.27	7.97	11.33
	NGDP	5	1.12	7.83	8.62	12.83
	OPT	2	0.82	8.94	9.87	14.55
First-stage test using $\mathcal{H}^*_{KLM}$	OG	5	0.07	9.69	10.72	15.65
	TR	12	3.15	6.94	7.63	11.23
	PEG	7	1.54	7.01	7.61	10.50
	NGDP	6	0.65	7.32	8.01	11.73
	OPT	4	0.20	8.05	8.92	13.07
Implementability test	OG	39	9.54		52.19	
	TR	42	8.94		55.76	
	PEG	39	9.67		52.19	
	NGDP	41	10.70		54.57	
	OPT	34	17.04		85.96	

Notes: Results of various pre-tests for the counterfactual rules considered. The first panel reports the test statistic and critical values (5% critical value for a bias tolerance of 10%) for the first-stage test for weak identification using the horizons $0, \ldots, \bar{H} - 1$ for each counterfactual rule, based on the Aruoba and Drechsel (2024) and Miranda-Agrippino and Ricco (2021) shocks. The next two panels report the optimal horizons, $\mathcal{H}^*_{AR}$ and $\mathcal{H}^*_{KLM}$, and the associated first-stage test statistics and critical values. The final two panels report the test statistics and (robust) 5% critical values for the test for implementability of each counterfactual rule.

but sometimes larger than using all horizons. While using our criterion to select $\mathcal{H}$ does not always reduce the evidence of weak identification bias in β, recall that our criterion is based on the volume of the robust confidence sets (and also the bias relative to the responses using all horizons), which may not be directly informative about the bias in an individual variable's counterfactual response. Indeed, a rule that selected horizons to maximise the first-stage test statistic would result in a substantial pre-testing problem, and invalidate subsequent inference.

The final panel of Table 3 report the test statistics and critical values for the identification-robust version of our test for implementability, considering all $\bar{H}$ horizons. For none of the counterfactuals are we able to reject the null hypothesis that the rule is implementable using the two shocks considered. The failures to reject are decisive enough that these results are unchanged even if we instead compare the test statistics to the smaller critical values that apply if strong identification

were assumed.[6]

Finally, we are ready to report the responses under each of the counterfactual rules. For each rule, we report a) the IRF to a technology shock in the data (dashed black line), b) the point estimate based on $\mathcal{H}^*_{KLM}$ (orange line), c) the point estimate using horizons $0, \ldots, \bar{H}$ (red line), d) the KLM 68% robust confidence set (shaded green), and e) the McKay and Wolf (2023) $16^{th} - 84^{th}$ percentile confidence bands. Figures 4-8 report these results for each of the counterfactual rules. Figures 14-28 in the Appendix report analogous results using the AR confidence set and the associated set of optimal horizons, $\mathcal{H}^*_{AR}$.

First, we note that for nearly all horizons for all counterfactuals, our frequentist estimates based on $\mathcal{H}^*_{KLM}$ or $0, \ldots, \bar{H} - 1$ lie within McKay and Wolf's (2023) confidence bands. Exceptions typically occur between $h = 3$ and $h = 6$, where there are kinks in several of the responses, which also appear in the underlying OLS IRFs. These are smoothed out to some extent by Bayesian estimation relative to OLS.

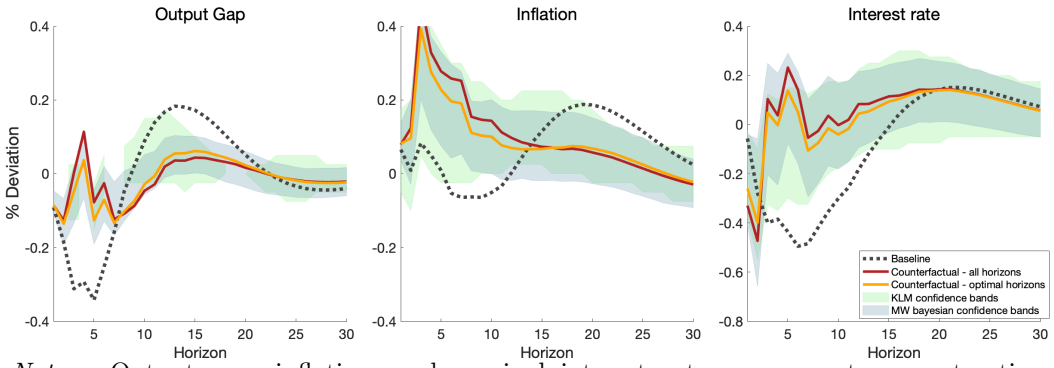
Next, note that for GAP, PEG, and OPT rules, we do not report confidence intervals for some variable-horizon combinations at all. This is a result of Remark 3, where we explain that, for rules that depend on a single variable (or on multiple variables in particular ways for OPT) – if the counterfactual rule is indeed implementable – the response must be zero, by assumption. As we explain, it is therefore not meaningful to conduct asymptotic inference for such responses.

We now examine the implications of our selection of $\mathcal{H}^*_{KLM}$. For TR, PEG, and OPT, the counterfactual responses based on $\mathcal{H}^*_{KLM}$ are almost exactly the same as those using $0, \ldots, \bar{H} - 1$. This means that our selection of a smaller number of horizons does not meaningfully impact our ability to impose the counterfactual rule at all periods, t. For NGDP, this is the case for y_t and π_t, but there are some small deviations from horizons 1 to 15 for i_t. However, given that the point of the exercise is to determine how a counterfactual monetary policy rule would impact economic outcomes, to the extent that the responses of those outcomes are essentially unchanged, the approximation error in the path of the interest rate can be seen as largely inconsequential. For GAP, there are small deviations at some medium-run horizons for all variables, although they are small relative to the size of the confidence intervals. Therefore, we conclude that restricting our attention to a smaller set of horizons for the purpose of strengthening identification results in little bias in the estimated counterfactual responses.

Finally, we turn to the economic conclusions that we can draw from the counterfactuals themselves, and compare them to the conclusions in McKay and Wolf (2023). In general, our KLM confidence sets based on $\mathcal{H}^*_{KLM}$ are somewhat wider than the Bayesian bands computed using their procedure and horizons, although

[6] Andrews et al. (2019) note that in practice there do not seem to be substantial size distortions from using the strong-identification J-test even when identification is weak.

Figure 4: Policy counterfactual, output gap targeting



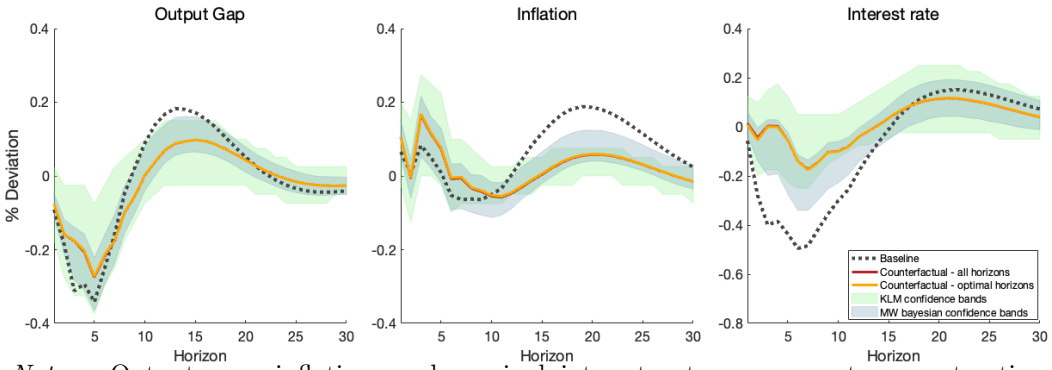
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to output gap targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to output gap targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

for some variables and horizons for some counterfactuals they are rather comparable.

Output gap targeting. For the output gap (Figure 4), we know that if the counterfactual rule is truly implemented, then the response must be zero, so the estimates in the first panel are informative only of the extent of approximation error, and not economically meaningful. To that end, zero is contained in the robust interval for nearly all horizons. As in McKay and Wolf (2023), leaning against the output decline resulting from the technology shock comes at the cost of significantly higher inflation at medium horizons, although we lose significance at shorter horizons compared to their estimates. Naturally, this change in behaviour is driven by a sharp decline in the interest rate on impact, although, unlike in the original McKay and Wolf (2023) specification, this is not statistically significantly lower than the response in the data. Note however, that this results from our use of a different VAR and monetary policy shocks than the original paper, as the Bayesian confidence bands we report based on our specification confirm. As in McKay and Wolf (2023), we do find a significantly higher interest rate in the short to medium run, as there is no prolonged and deep contraction for the central bank to respond to.

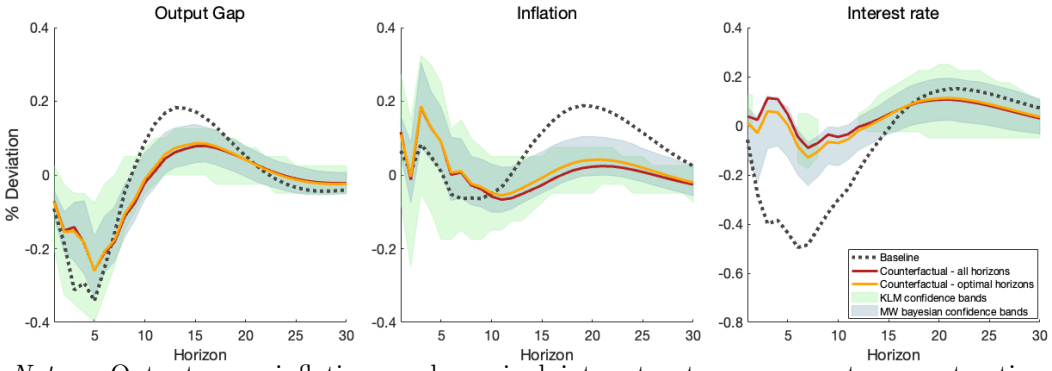
Taylor Rule. Our results for the Taylor Rule (Figure 5) match those in McKay and Wolf (2023) well. The response of the output gap is just significantly lower at medium horizons. The inflationary effect of the shock is significantly attenuated at medium and long horizons. This change is driven by a significantly less-accommodative nominal interest rate at short and medium horizons.

Figure 5: Policy counterfactual, Taylor Rule



Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to a simple Taylor Rule using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to the same Taylor Rule using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 6: Policy counterfactual, interest rate peg

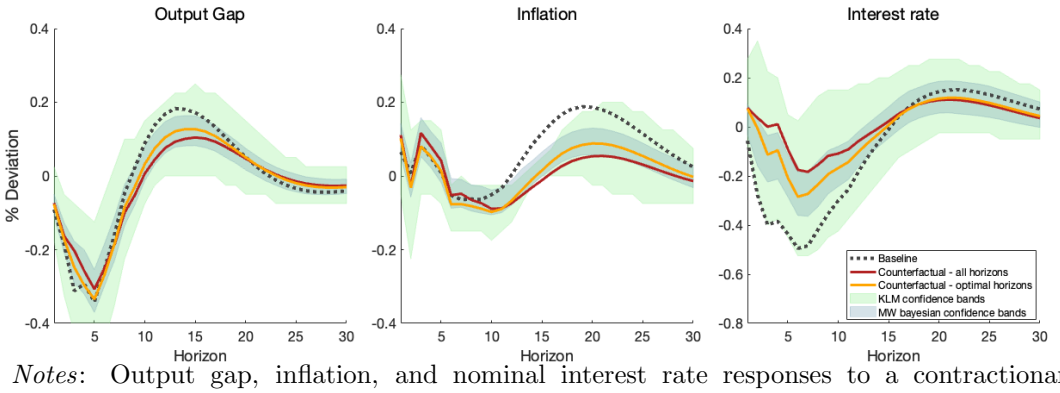


Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to interest rate targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to interest rate targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Interest rate peg. For the interest rate peg (Figure 6), we find that that the output gap rises significantly less in the medium run as it rebounds following the initial contraction. This conclusion is shared by the McKay and Wolf (2023) procedure in our specification, although significance is rather borderline in their paper. Turning to inflation, we share the conclusion that the counterfactual response is significantly dampened in the medium- and long-run. Any deviation of the interest rate response from zero is evidence of an approximation error; our robust band contains zero for all plotted horizons, but this is not the case for the Bayesian band.

Nominal GDP targeting. For nominal GDP targeting (Figure 27), unlike McKay and Wolf (2023), we do not find that the marginally lower response of the output gap is significant in the short- to medium-run. However, we share the

Figure 7: Policy counterfactual, nominal GDP targeting



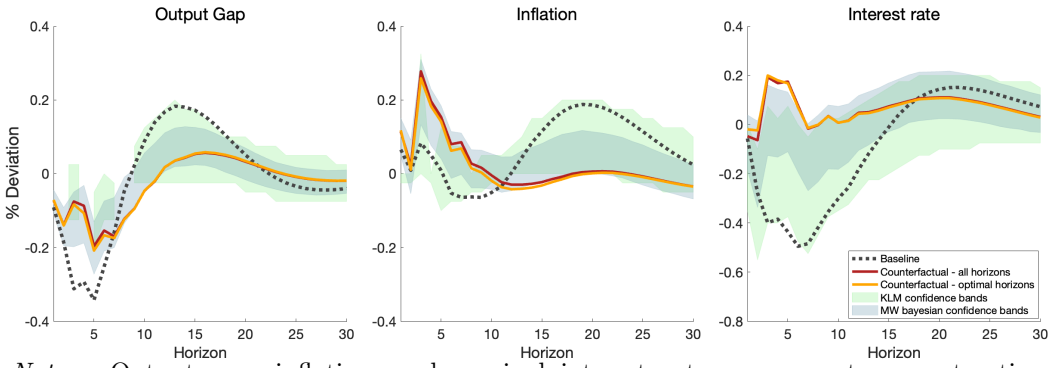
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to nominal GDP targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to nominal GDP targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

finding of significantly lower inflation in the medium-run. This is driven by a smaller cut in the nominal interest rate in the short- to medium-run, with the difference considerably larger than in McKay and Wolf's (2023) reported results (although only on the edge of statistical significance).

Optimal Average inflation targeting. Under an optimal average inflation targeting monetary policy rule (Figure 8), like McKay and Wolf (2023), we find a significantly smaller contraction in output in the short-run and a (marginally significant) lower response at some medium-run horizons. For inflation, we likewise share the finding of somewhat higher values at some short- to medium-run horizons. Optimal policy dictates a dramatically different path for the nominal interest rate (rising at horizons around a year), although these values are not actually contained in the robust confidence interval. Based on the interval, we conclude that the interest rate is sigificantly less accomodative under optimal policy, although only for a much smaller range of horizons than McKay and Wolf (2023).

This empirical illustration highlights several of the main points to emerge from our framework for inference. First, identification does indeed prove to be quite weak in practice, in line with our interpretation of the estimator as a weak instruments problem. This is borne out both by our first-stage test statistic and the width of the robust confidence sets that we report. In Appendix A.3, we report standard Wald intervals for each counterfactual for comparison, which further highlights the scope of weak identification – these intervals are considerably narrower than the robust intervals plotted in Appendix A.1 (both sets use the maximum feasible number of horizons). The policy rules appear implementable through the lens of our simple test. Our selection rule for $\mathcal{H}$ significantly sharpens inference at the cost of minimal bias in the counterfactual responses for nearly all variables and rules. Following a point made throughout the paper (see, e.g., Remark 3), we opt to not interpret

Figure 8: Policy counterfactual, Optimal average inflation targeting



Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to an optimal average inflation targeting monetary policy rule using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to the same optimal average inflation targeting monetary policy rule using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

the counterfactual paths for some variables and counterfactuals, since they do not represent meaningful targets for inference, once implementability is assumed.

Empirically, our results show that – using our tailored set of tools – it is possible to arrive at broadly similar conclusions to those of McKay and Wolf (2023) even after weak identification is taken into account and robust inference is conducted. While the differences between counterfactual and observed responses may not be significant at quite as many horizons using our robust methods, the main conclusions prevail. Thus, acknowledging the challenges to identification present in the methodology should not be seen as prohibitive to obtaining economically meaningful empirical results.

6 Conclusion

We develop a frequentist framework for estimation and inference of macroeconomic counterfactuals, exploiting the identification argument in McKay and Wolf (2023). This frequentist perspective makes clear that weak identification is likely to be a key challenge, as it indeed proves to be in our empirical analysis. As a result, we provide researchers with disciplined guidance on choosing the horizons used for estimation, a key ingredient to the estimator that is closely linked to the strength of identification. Combined with pre-tests for both weak identification and the implementability of a counterfactual rule, we provide researchers with valuable tools to craft appropriate specifications to estimate macroeconomic counterfactuals. Our proposed inference techniques are straightforward to apply whether identification is deemed strong or weak.

Empirical macroeconomics – and in particular the identification of IRFs – has recently begun to confront the fact that identification is likely weak in many set-

tings. The counterfactual exercise of McKay and Wolf (2023) – while clever – asks much more of the data than conventional analysis of observed policies. Awareness of weak identification and how to overcome its threats is essential for the credibility of the emerging body of empirical work based on their approach. The most obvious avenue for future work is to apply our methods to alternative empirical settings, in particular fiscal policy. With formal tools to assess identification strength, it is now also possible to quantitatively evaluate the identifying variation available in different combinations of the many available monetary policy shocks. Theoretically, as we highlighted in the introduction, developing a fully formal Bayesian procedure, under which the researcher imposes priors on the counterfactual response or β and samples from their posteriors directly would offer greater transparency and be appealing in practice, since it would offer a natural way to discipline the model (through prior information) and potentially avoid the symptoms of weak identification.

References

Anderson, T. W., & Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, *20*(1), 46 –63.

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*(1), 727–753.

Aruoba, S. B., & Drechsel, T. (2024). *Identifying monetary policy shocks: A natural language approach* (Working Paper No. 32417). National Bureau of Economic Research. https://doi.org/10.3386/w32417

Barnichon, R., & Mesters, G. (2020). Identifying modern macro equations with old shocks. *The Quarterly Journal of Economics*, *135*(4), 2255–2298.

Barnichon, R., & Mesters, G. (2023). A sufficient statistics approach for macro policy. *American Economic Review*, *113*(11), 2809–45. https://doi.org/10.1257/aer.20220581

Ben Zeev, N., & Khan, H. (2015). Investment-specific news shocks and us business cycles. *Journal of Money, Credit and Banking*, *47*(7), 1443–1464.

Galí, J. (2015). *Monetary policy, inflation, and the business cycle: An introduction to the new keynesian framework and its applications*. Princeton University Press.

Gertler, M., & Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, *7*(1), 44–76. https://doi.org/10.1257/mac.20130329

Giannone, D., Lenza, M., & Primiceri, G. (2025). *Bayesian inference in iv regressions* (tech. rep.).

Hall, A. R., Inoue, A., Nason, J. M., & Rossi, B. (2012). Information criteria for impulse response function matching estimation of dsge models [Thirtieth Anniversary of Generalized Method of Moments]. *Journal of Econometrics*, *170*(2), 499–518.

Kleibergen, F. (2005). Testing Parameters in GMM Without Assuming that They Are Identified. *Econometrica*, *73*(4), 1103–1123.

Kleijn, B. J. K., & van der Vaart, A. W. (2012). The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, *6*. https://doi.org/10.1214/12-EJS675

Lewis, D. J., & Mertens, K. (2025a). A robust test for weak instruments for 2sls with multiple endogneous regressors. *Review of Economics Studies*.

Lewis, D. J., & Mertens, K. (2025b). Dynamic identification using system projections and instrumental variables.

Lewis, D. J., & Mertens, K. (2026). *Weak instrument bias in impulse response estimators* (Working Papers No. CEPR Discussion Paper No. 20990). CEPR. https://cepr.org/publications/dp20990

Lucas, R. E. Econometric policy evaluation: A critique. In: In *Carnegie-rochester conference series on public policy. 1*. North-Holland. 1976, 19–46.

McKay, A., & Wolf, C. K. (2023). What can time-series regressions tell us about policy counterfactuals? *Econometrica*, *91*(5), 1695–1725.

Miranda-Agrippino, S., & Ricco, G. (2021). The transmission of monetary policy shocks. *American Economic Journal: Macroeconomics*, *13*(3), 74–107. https://doi.org/10.1257/mac.20180124

Montiel Olea, J. L., & Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, *31*(3), 358–369.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley; Sons Ltd.

Nelson, C. R., & Startz, R. (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business*, *63*(1), S125–S140.

Plagborg-Møller, M., & Wolf, C. K. (2021). Local Projections and VARs Estimate the Same Impulse Responses. *Econometrica*, *89*(2), 955–980.

Ramey, V. (2016). Macroeconomic Shocks and Their Propagation. In J. B. Taylor & H. Uhlig (Eds.). Elsevier.

Romer, C., & Romer, D. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, *94*(4), 1055–1084. https://doi.org/10.1257/0002828042002651

Sims, C. A., & Zha, T. (2006). Does monetary policy generate recessions? *Macroeconomic Dynamics*, *10*(2), 231–272.

Stock, J., & Yogo, M. (2005). Testing for weak instruments in linear iv regression. In D. W. Andrews (Ed.), *Identification and Inference for Econometric Models* (pp. 80–108). Cambridge University Press.

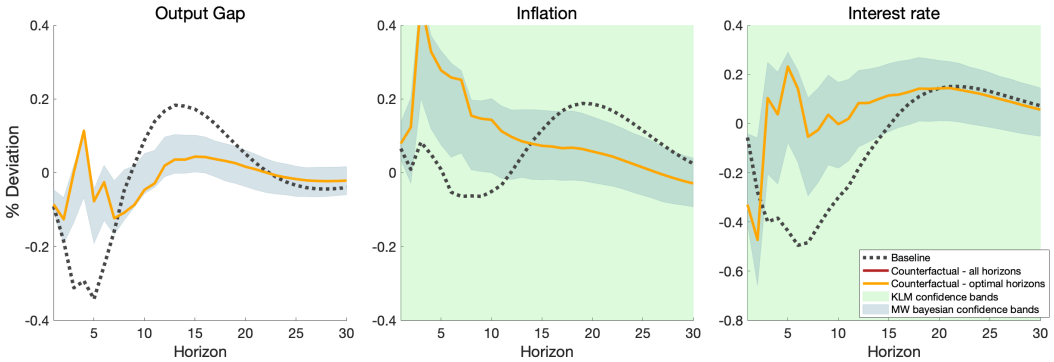
Stock, J. H., & Wright, J. H. (2000). GMM with Weak Identification. *Econometrica*, *68*(5), 1055–1096.

A Additional empirical results

A.1 Counterfactuals using $\mathcal{H} = 0, \ldots, \bar{H} - 1$

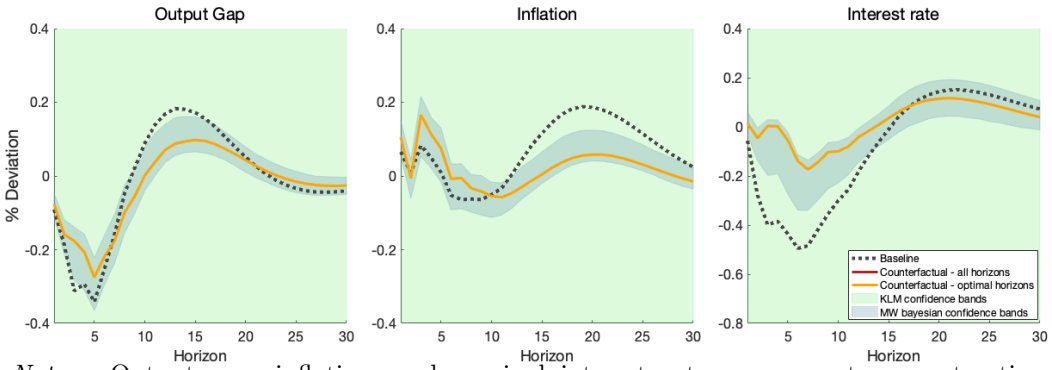
Below we report counterfactuals estimated using all feasible horiozns $\mathcal{H} = \{0, \ldots, \bar{H} - 1\}$. Note that for variables entering a single-variable policy rule (GAP, PEG), we report no robust confidence bands since the response is zero by assumption at all horizons if the rule is implementable. For all variable-rule combinations except for inflation for NGDP, the robust confidence bands cover the entire plotted space; meaningful economic conclusions are not possible.

Figure 9: Policy counterfactual, output gap targeting



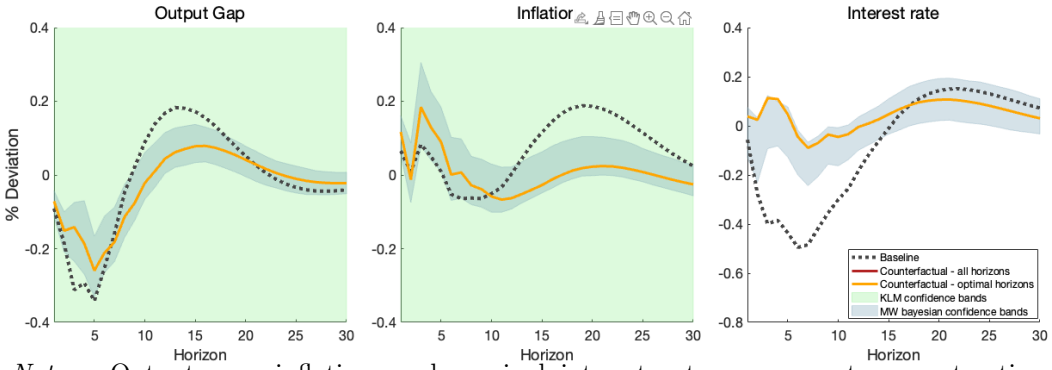
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line) and the best feasible approximation to output gap targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 10: Policy counterfactual, Taylor Rule



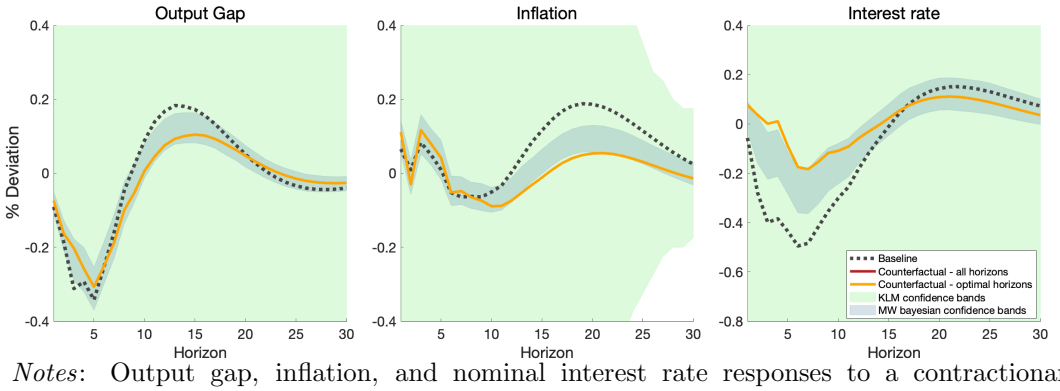
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line) and the best feasible approximation to the same Taylor Rule using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 11: Policy counterfactual, interest rate peg



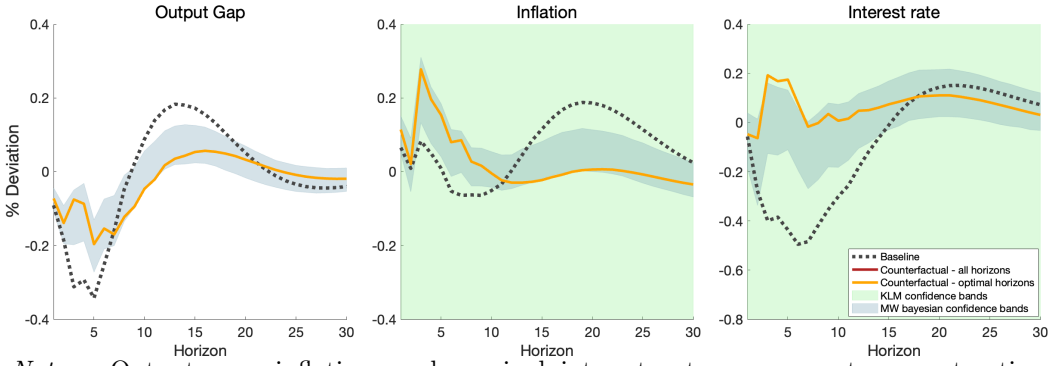
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line) and the best feasible approximation to interest rate targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 12: Policy counterfactual, nominal GDP targeting



Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line) and the best feasible approximation to nominal GDP targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 13: Policy counterfactual, Optimal average inflation targeting

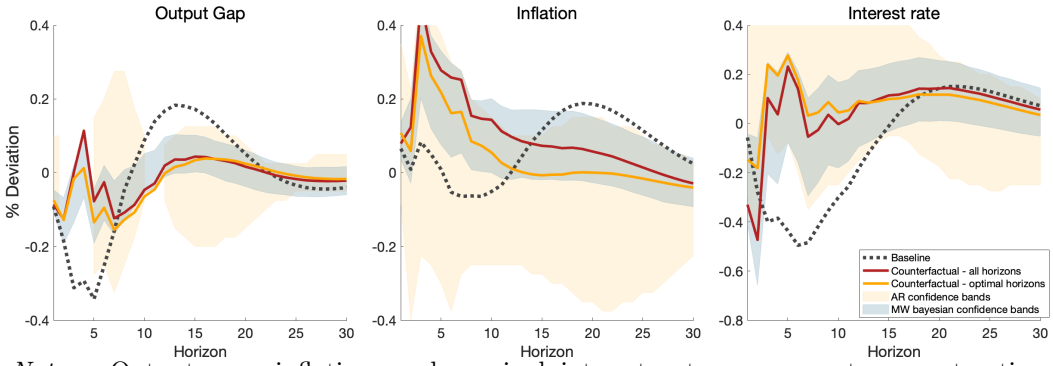


Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line) and the best feasible approximation to the same optimal average inflation targeting monetary policy rule using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

A.2 Counterfactuals using the AR confidence set

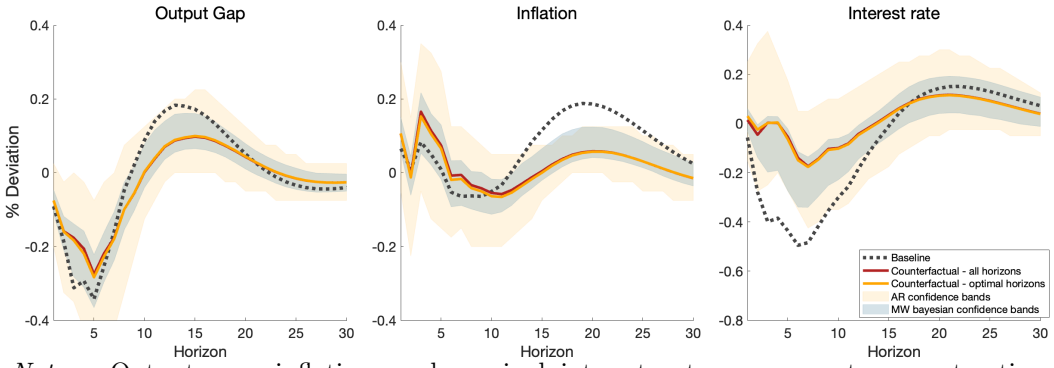
Below, we repeat our analysis using Anderson and Rubin (1949) confidence bands and associated optimal horizons, $\mathcal{H}^*$, instead of the KLM used in the text. Although these bands are sometimes wider, (since all specifications are over-identified), many of the economic conclusions in our main results are replicated, establishing the robustness of our findings.

Figure 14: Policy counterfactual, output gap targeting



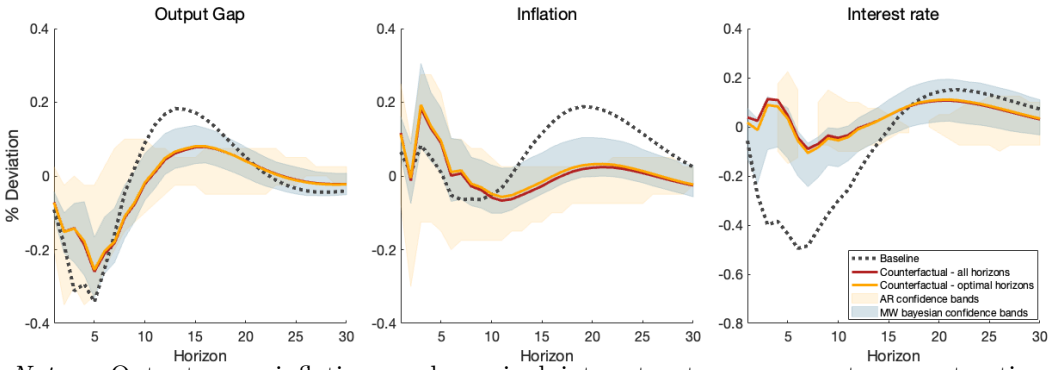
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to output gap targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to output gap targeting using horizons $0 : \bar{H} - 1$. The yellow shaded area is the AR 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 15: Policy counterfactual, Taylor Rule



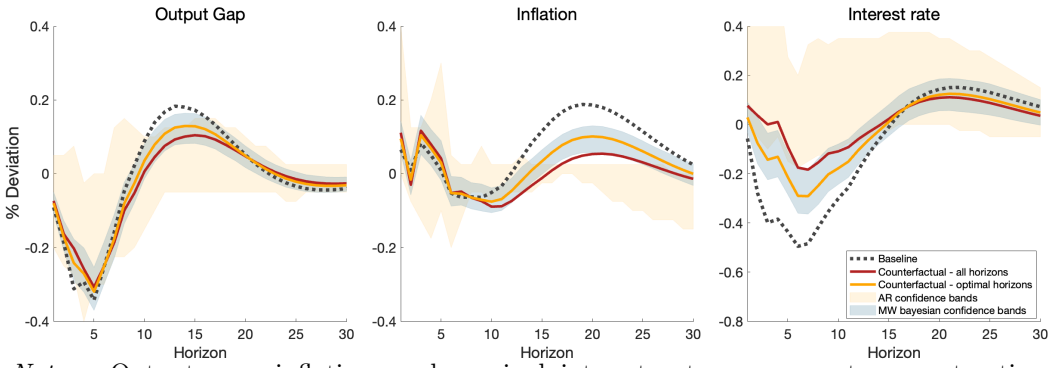
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to a simple Taylor Rule using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to the same Taylor Rule using horizons $0 : \bar{H} - 1$. The green shaded area is the AR 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 16: Policy counterfactual, interest rate peg



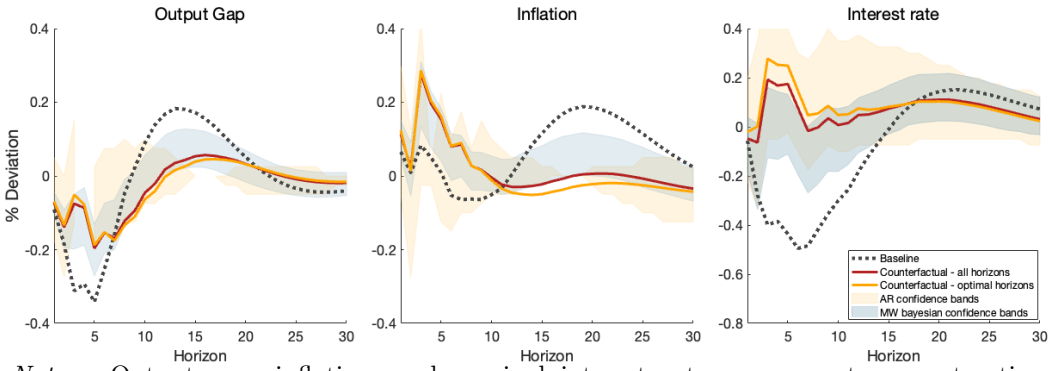
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to interest rate targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to interest rate targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the AR 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 17: Policy counterfactual, nominal GDP targeting



Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to nominal GDP targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to nominal GDP targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the AR 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 18: Policy counterfactual, Optimal average inflation targeting

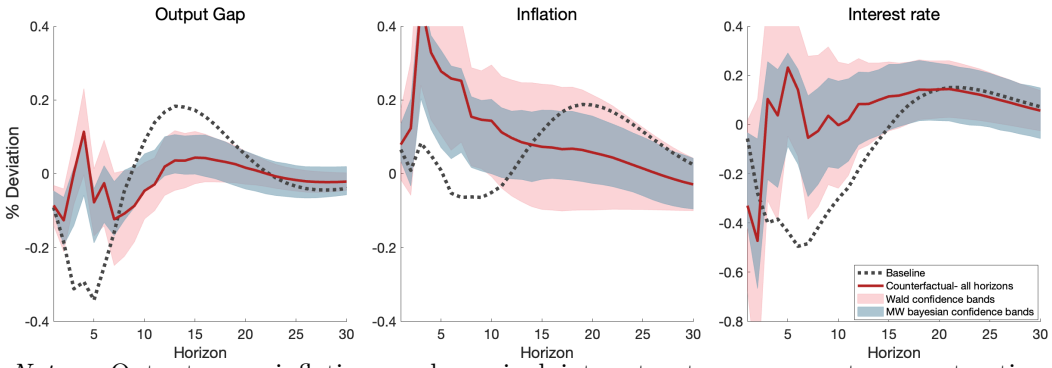


Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to an optimal average inflation targeting monetary policy rule using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to the same optimal average inflation targeting monetary policy rule using horizons $0 : \bar{H} - 1$. The green shaded area is the AR 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

A.3 Counterfactuals using the Wald confidence set

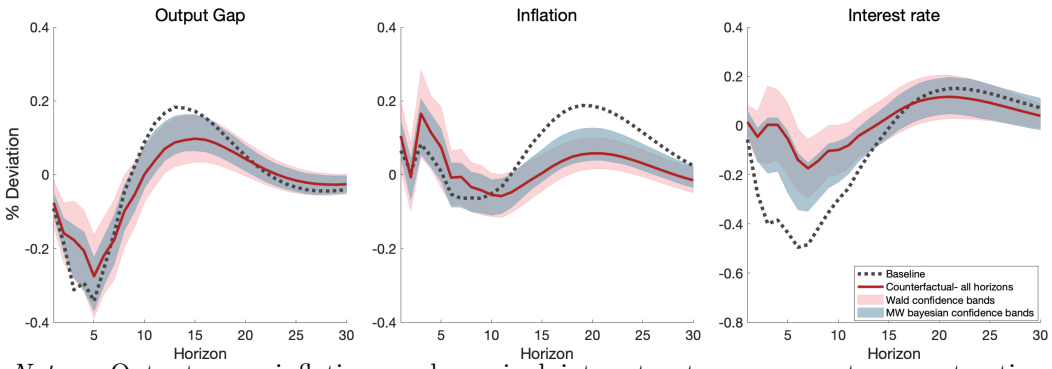
In this section we plot estimates based on all feasible horiozns $\mathcal{H} = \{0, \ldots, \bar{H} - 1\}$ with Wald confidence bands. These are not robust to weak identification. These bands are generally considerably narrower than the robust bands reported in the text and typically a similar width to the Bayesian ones. As a result, these estimates lead to much stronger conclusions than warranted for many variables and counterfactual rules.

Figure 19: Policy counterfactual, output gap targeting



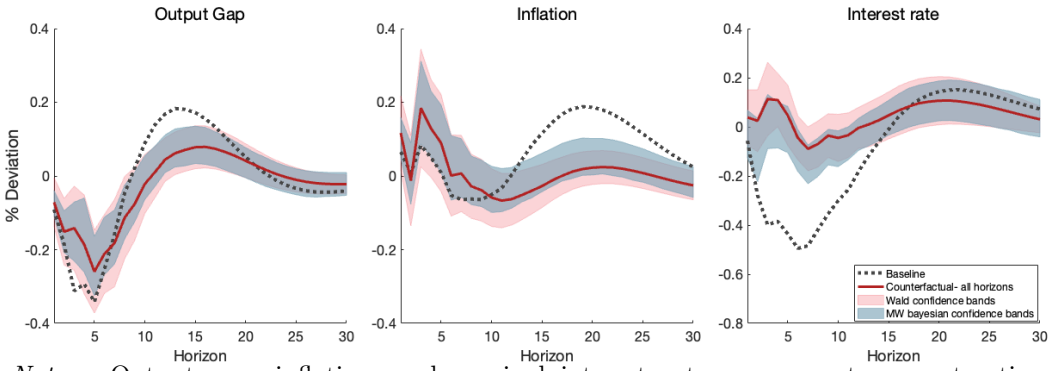
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to output gap targeting using horizons $\mathcal{H}^*$ (orange line, chosen based on KLM), and the best feasible approximation to output gap targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the Wald 68% confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 20: Policy counterfactual, Taylor Rule



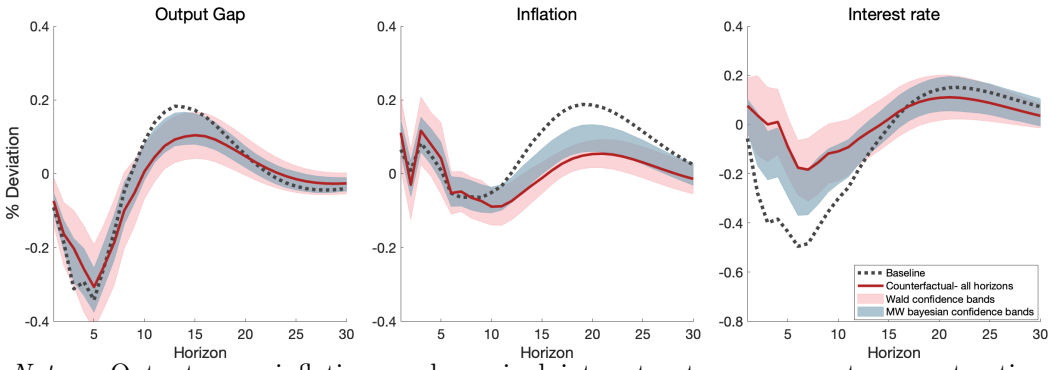
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to a simple Taylor Rule using horizons $\mathcal{H}^*$ (orange line, chosen based on KLM), and the best feasible approximation to the same Taylor Rule using horizons $0 : \bar{H} - 1$. The green shaded area is the Wald 68% confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 21: Policy counterfactual, interest rate peg



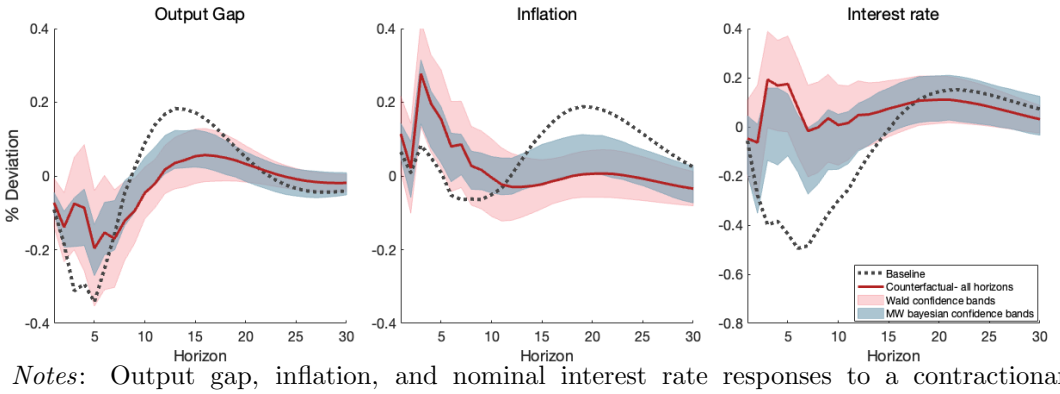
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to interest rate targeting using horizons $\mathcal{H}^*$ (orange line, chosen based on KLM), and the best feasible approximation to interest rate targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the Wald 68% confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 22: Policy counterfactual, nominal GDP targeting



Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to nominal GDP targeting using horizons $\mathcal{H}^*$ (orange line, chosen based on KLM), and the best feasible approximation to nominal GDP targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the Wald 68% confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 23: Policy counterfactual, Optimal average inflation targeting

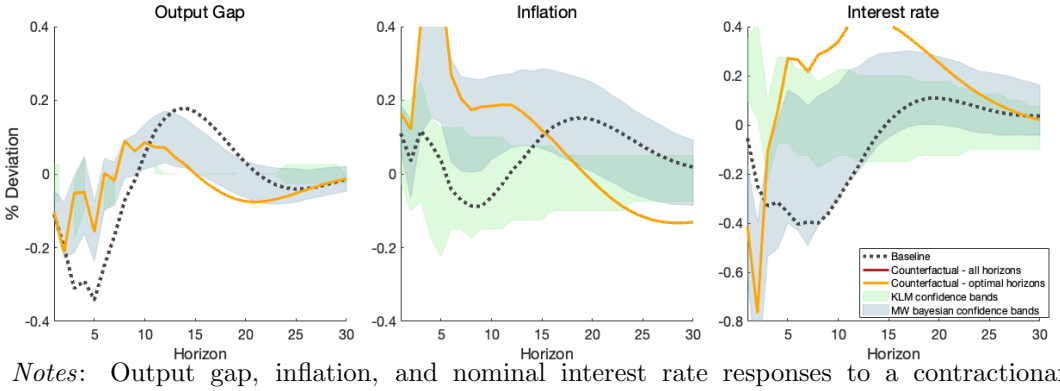


Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to an optimal average inflation targeting monetary policy rule using horizons $\mathcal{H}^*$ (orange line, chosen based on KLM), and the best feasible approximation to the same optimal average inflation targeting monetary policy rule using horizons $0 : \bar{H} - 1$. The green shaded area is the Wald 68% confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

A.4 Counterfactuals using Romer and Romer (2004) and Gertler and Karadi (2015) shocks

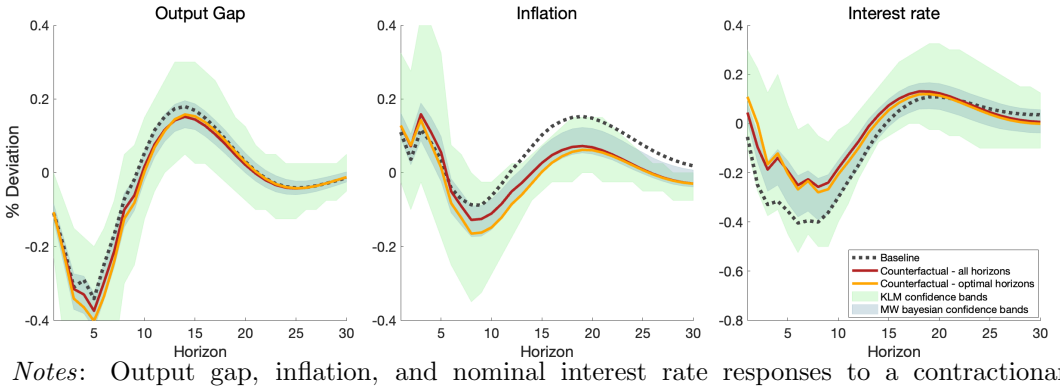
In this section, we report results using the Romer and Romer (2004) and Gertler and Karadi (2015) shocks, as in McKay and Wolf's (2023) baseline. As noted in the text, these shocks generally lead to weaker identification; this is most clear for NGDP. However, for several-rule pairs for which the counterfactual responses are significantly different from the observed responses in the main text, the difference is no longer significant using these alternative shocks. For GAP, the bands do not tend to contain the point estimates, which makes interpretation difficult.

Figure 24: Policy counterfactual, output gap targeting



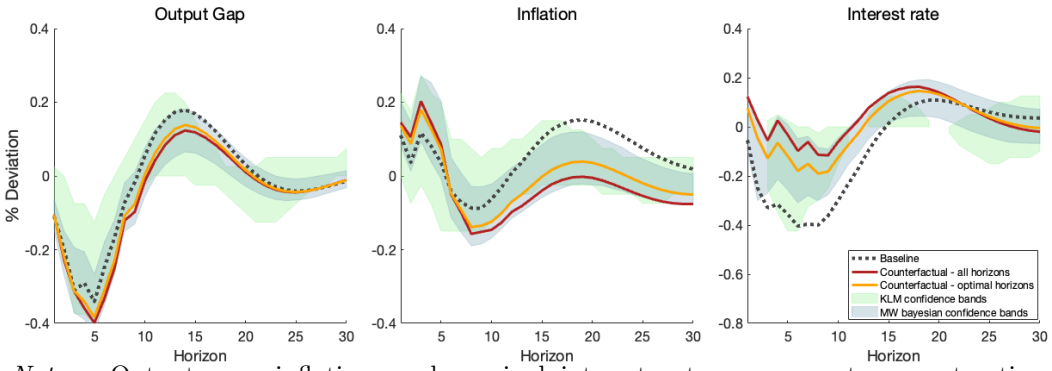
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to output gap targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to output gap targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 25: Policy counterfactual, Taylor Rule



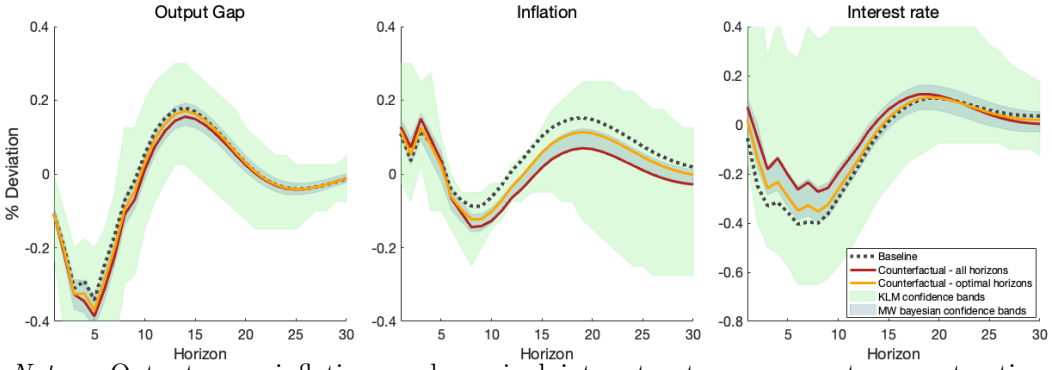
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to a simple Taylor Rule using horizons $\mathcal{H}^*$ (orange line, chosen based on KLM), and the best feasible approximation to the same Taylor Rule using horizons $0 : \bar{H} - 1$. The green shaded area is the Wald 68% confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 26: Policy counterfactual, interest rate peg



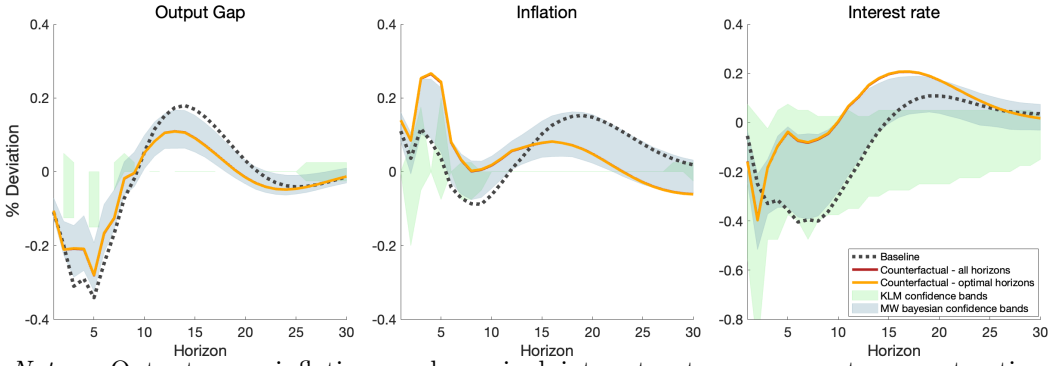
Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to interest rate targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to interest rate targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 27: Policy counterfactual, nominal GDP targeting



Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to nominal GDP targeting using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to nominal GDP targeting using horizons $0 : \bar{H} - 1$. The green shaded area is the KLM 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

Figure 28: Policy counterfactual, Optimal average inflation targeting



Notes: Output gap, inflation, and nominal interest rate responses to a contractionary investment-specific technology shock under the prevailing rule (dashed black line), the best feasible approximation to an optimal average inflation targeting monetary policy rule using horizons $\mathcal{H}^*$ (orange line), and the best feasible approximation to the same optimal average inflation targeting monetary policy rule using horizons $0 : \bar{H} - 1$. The green shaded area is the AR 68% robust confidence set and the blue shaded area is the McKay and Wolf (2023) Bayesian $16^{th} - 84^{th}$ percentile confidence band.

B Additional simulation results

In this section, we report results for simulations for an alternative calibration of the model in Section 4.5 with stronger instruments. In particular, we set the volatility of the monetary policy news shocks to be equal to 1 instead of 0.5, so that it is the same as that of the other shocks in the model, and the shocks explain more variation in observables. Table 4 reports the results.

As expected, the bias is considerably smaller for both models than for the weaker calibration in the main text. The RMSEs are mostly lower as well. The biggest difference is the coverage of Wald inference. In the $K = 5$ model, coverage is now around the nominal level jointly across horizons, but higher for individual horizons, tending towards 1 for longer horizons. This shows that even though there may no longer be notable positive size distortions, the normal distribution still does not provide a good approximation of the distribution of the t-statistics for counterfactual estimates. However, for $K = 2$, coverage remains low, especially for β (0.76), jointly across counterfactual horizons (0.86), and for $h = 0$ (0.83). The robust inference methods offer good coverage, close to the nominal level for all models and targets, even when the horizons are chosen optimally. This is the case even though there is now substantial variation in the horizons selected ($\overline{|\mathcal{H}^*_{KLM}|} = 3.60$ for $T = 500$).

Turning to test statistics, identification remains apparently weak according to the bias-based test in the $K = 5$ model. However, identification is now detectably stronger, with average test statistics of 1.00 ($T = 150$) and 2.92 ($T = 500$), compared to critical values of around 3.5, on average. When H is chosen optimally, the average test statistic rises to 4.29 for $T = 500$, with an average critical value of 4.20, corresponding to a rejection rate of 0.48, demonstrating non-trivial power. The empirical size of our test for implementability using strong identification critical values now approaches its nominal level for larger T, confirming the observation of Andrews et al. (2019) that frequently using such values does not lead to over-rejection, even under strong identification. The test remains conservative using weak identification critical values.

Table 4: Performance in simulations: stronger instruments

$H=5, K=5$					$T=150$							$T=500$		
	β	Joint	$h=0$	1	2	3	4	β	Joint	$h=0$	1	2	3	4
Bias	-0.08	1.58	0.34	1.64	2.37	3.49	0.07	1.47	3.99	1.33	3.20	3.13	6.84	5.46
RMSE	104.97	120.08	93.66	125.08	129.51	166.10	86.07	250.88	243.72	192.09	271.57	139.47	323.91	291.55
Wald coverage	0.94	0.95	0.96	0.97	0.98	1.00	1.00	0.96	0.96	0.97	0.97	0.98	0.99	1.00
AR/KLM coverage	0.94	0.92	0.92	0.94	0.94	0.94	0.94	0.95	0.94	0.94	0.95	0.95	0.95	0.95
Test statistics			$g_{min}^{avg}=0.03$		$cv=8.25$					$g_{min}^{avg}=0.17$		$cv=6.53$		
$H=5, K=2$					$T=150$							$T=500$		
	β	Joint	$h=0$	1	2	3	4	β	Joint	$h=0$	1	2	3	4
Bias	0.56	0.16	0.48	0.24	0.05	0.02	0.01	0.18	0.05	0.16	0.06	0.01	0.02	0.02
RMSE	1.01	0.75	0.91	0.84	0.68	0.66	0.65	0.54	0.43	0.46	0.49	0.40	0.39	0.40
Wald coverage	0.76	0.86	0.83	0.95	0.99	0.99	0.99	0.88	0.91	0.90	0.96	0.97	0.97	0.97
AR coverage	0.95	0.95	0.94	0.94	0.95	0.95	0.95	0.95	0.94	0.94	0.95	0.95	0.95	0.95
KLM coverage	0.93	0.94	0.92	0.93	0.94	0.94	0.94	0.95	0.94	0.95	0.95	0.94	0.95	0.95
Test statistics	$g_{min}^{avg}=1.00$		$cv=3.63$		$\hat{\alpha}_s(J)=0.01$		$\hat{\alpha}_w(J)=0.00$	$g_{min}^{avg}=2.92$		$cv=3.35$		$\hat{\alpha}_s(J)=0.04$		$\hat{\alpha}_w(J)=0.01$
$H=\lvert\mathcal{H}_{KLM}^*\rvert, K=2$				$T=150, \overline{\lvert\mathcal{H}_{KLM}^*\rvert}=4.54$							$T=500, \overline{\lvert\mathcal{H}_{KLM}^*\rvert}=3.60$			
	β	Joint	$h=0$	1	2	3	4	β	Joint	$h=0$	1	2	3	4
Bias (true)	0.53	0.17	0.47	0.25	0.06	0.04	0.03	0.08	0.03	0.07	0.03	0.01	0.01	0.02
RMSE (true)	1.12	0.81	0.97	0.93	0.73	0.72	0.71	0.72	0.48	0.57	0.55	0.45	0.43	0.42
AR coverage (true)	0.95	0.94	0.94	0.94	0.95	0.95	0.95	0.96	0.96	0.95	0.95	0.96	0.96	0.96
KLM coverage (true)	0.94	0.93	0.93	0.93	0.94	0.94	0.95	0.93	0.94	0.93	0.94	0.94	0.94	0.4
Test statistics	$g_{min}^{avg}=1.17$		$cv=3.99$		$\hat{\alpha}_s(J)=0.01$		$\hat{\alpha}_w(J)=0.00$	$g_{min}^{avg}=4.29$		$cv=4.20$		$\hat{\alpha}_s(J)=0.04$		$\hat{\alpha}_w(J)=0.01$

Notes: The table reports simulation performance for the three-equation NK model described in the text with stronger monetary policy news instruments, for a Taylor Rule counterfactual output response. The top panel considers a model with $K=5$ monetary policy shocks, all of which are available to the econometrician, while in the middle and bottom, only two shocks are available, but these suffice to implement the counterfactual rule. For the top two panels, five horizons are used for estimation, and for the bottom, horizons are chosen to minimise (55). The left panels consider $T=150$, and the right $T=500$. For each panel, the top row reports the bias for β, averaged across horizons, and by horizon. The second row reports the RMSE for β, averaged across horizons, and by horizon. The next rows report Wald, AR, and KLM coverage for nominal 95% intervals. Under "Test statistics", we report the average value of g_{min} and critical values for weak identification as well as the empirical rejection rates of our test for implementability using critical values valid under both strong and weak identification ($\hat{\alpha}_s(J), \hat{\alpha}_w(J)$, respectively).

C Expressions for inference

C.1 Computation of V_β

The asymptotic variance of $\hat{\beta}$ in Proposition 2 is $V_\beta = D_\beta \mathbf{M} D_\beta'$, where

$$D_\beta = -\begin{bmatrix}(\Phi'\Phi)^{-1}\Phi' & L_\beta\end{bmatrix} D_{\Psi\Phi}, \tag{59}$$

and

$$L_\beta = \Psi' \left[I_H \otimes (\Phi'\Phi)^{-1}\mathcal{K}_{H,K} + (\Phi \otimes I_K)(-(\Phi'\Phi)^{-1} \otimes (\Phi'\Phi)^{-1}) \right. \tag{60}$$

$$\left. \times (I_K \otimes \Phi' + (\Phi' \otimes I_K)\mathcal{K}_{H,K}) \right] \tag{61}$$

and $\mathcal{K}_{H,K}$ represents the commutation matrix, so that $\text{vec}(\Phi') = \mathcal{K}_{H,K}\,\text{vec}(\Phi)$.

C.2 Computation of $KLM(\beta)$

The KLM statistic for β is defined in the text as

$$KLM(\beta) = T^{-1}\tilde{q}(\beta)\left[\hat{\mathcal{D}}(\beta)'\hat{V}_f(\beta)^{-1}\hat{\mathcal{D}}(\beta)\right]^{-1}\tilde{q}(\beta)'. \tag{62}$$

Let $q_i = \hat{\Phi}_i$, where Φ_i is the i^{th} column of $\hat{\Phi}$. Then

$$\hat{\mathcal{D}}(\beta) = \left[q_1 - \hat{V}_{qf,1}(\beta)\hat{V}_f(\beta)^{-1}f(\beta), \ldots, q_K - \hat{V}_{qf,K}(\beta)\hat{V}_f(\beta)^{-1}f(\beta)\right], \tag{63}$$

where

$$V_{qf}(\beta) \equiv \left[V_{qf,1}(\beta)', \ldots, V_{qf,K}(\beta)'\right]' = [\hat{\mathbf{M}}_{21}\ \hat{\mathbf{M}}_2]D'_{\Psi\Phi}\left[1\ -\beta'\right]'. \tag{64}$$

C.3 Computation of V_Ξ

Let c denote the variable of interest. The asymptotic variance of $\hat{\Xi}^h$ in Proposition 6 is $V_\Xi = D_\Xi \mathbf{M} D'_\Xi$, where

$$D_\Xi = \left[\left(\mathbf{1}\{c = x_i\}e'_{i,n_x} \quad \mathbf{1}\{c = z_i\}e'_{i,n_z} \quad \beta' \otimes \left(\mathbf{1}\{c = x_i\}e'_{i,n_x} \quad \mathbf{1}\{c = z_i\}e'_{i,n_z}\right)\right) \otimes e'_{h,H}\right] + \Omega^h_c D_\beta \tag{65}$$

C.4 Computation of $V_g(\beta, \Xi^h)$

The asymptotic variance of $g(\beta, \Xi^h)$ in $AR(\beta, \Xi^h)$ is $V_g(\beta, \Xi^h) = D^h_{AR}\mathbf{M}D^{h\prime}_{AR}$, where

$$D^h_{AR} = \begin{bmatrix} D_{\Psi\Phi} \\ -\left(\mathbf{1}\{c = x_i\}e'_{i,n_x} \quad \mathbf{1}\{c = z_i\}e'_{i,n_z} \quad \beta' \otimes \left(\mathbf{1}\{c = x_i\}e'_{i,n_x} \quad \mathbf{1}\{c = z_i\}e'_{i,n_z}\right)\right) \otimes e'_{h,H} \end{bmatrix}. \tag{66}$$

C.5 Computation of $KLM(\beta, \Xi^h)$

The KLM statistic for $(\beta', \Xi_h)'$ is defined in the text as

$$KLM(\beta, \Xi^h) = T^{-1}\tilde{q}(\beta, \Xi)\left[\hat{\mathcal{D}}(\beta, \Xi^h)'\hat{V}_g(\beta, \Xi^h)^{-1}\hat{\mathcal{D}},(\beta, \Xi^h)\right]^{-1}\tilde{q}(\beta, \Xi^h)'. \tag{67}$$

where

$$\hat{\mathcal{D}}(\beta, \Xi^h) = \left[q_1 - \hat{V}_{qg,1}(\beta, \Xi^h)\hat{V}_g(\beta, \Xi^h)^{-1}g(\beta, \Xi^h), \ldots, q_{K+1} - \hat{V}_{qg,K+1}(\beta, \Xi^h)\hat{V}_g(\beta, \Xi^h)^{-1}g(\beta, \Xi^h)\right], \tag{68}$$

$q_j = (\Phi'_{\cdot,j},\ \Omega^h_{s,j,A})'$ $j = 1, \ldots, K$, $q_{K+1} = e_{H+1,H+1}$, $\Omega^h_{s,j,A}$ denotes the response to the j^{th} shock, and

$$V_{qg}(\beta, \Xi^h) \equiv \left[V_{qg,1}(\beta, \Xi^h)', \ldots, V_{qg,K+1}(\beta, \Xi^h)'\right]' = D^h_{KLM}\mathbf{M}D^{h\prime}_{AR} \tag{69}$$

where $D^h_{KLM} \equiv \left[D^{h\prime}_{KLM,1},\ \ldots,\ D^{h\prime}_{KLM,K},\ \mathbf{0}_{(H(n_z+n_x)(K+1))\times(H+1)}\right]'$ and

$$D^h_{KLM,j} = \begin{bmatrix} (\mathbf{0}_{H\times H} & e'_{j,K} \otimes I_H) & \times D_{\Psi\Phi} \\ \mathbf{0}_{1\times(H(n_x+n_z))} & \mathbf{1}\{c = x_i\}e'_{j,K} \otimes e'_{i,n_x} \otimes e'_{h,H} & \mathbf{1}\{c = z_i\}e'_{j,K} \otimes e'_{i,n_x} \otimes e'_{h,H} \end{bmatrix}. \tag{70}$$

D Test for weak identification

In this section, we present the theoretical results underlying the first stage test for weak identification.

First, we define what it means for identification to be weak.

Definition 1. *The weak identification set for is*

$$\mathbb{B}_\tau(\mathbf{W}) = \{C \in \mathbb{R}^{K\times H}, \beta \in \mathbb{R}^K : B_n > \tau\},$$

where τ is the bias tolerance (conventionally set to $\tau = 0.10$, following Stock and Yogo (2005)) and B_n is the Nagar approximation (described below) to B. Note that the weak identification set depends on $\mathbf{W}$, but also on $K \times H + K$ nuisance parameters that cannot be consistently estimated. We consider the worst-case over these unknown parameters to develop a procedure that controls size over all possible models.

The bias criterion in (34) has a convenient representation:

Lemma 1. *Under Assumptions 1 & 2, the bias criterion in Definition 34 can be decomposed as* $B = ||(\Delta^{-\frac{1}{2}}\Omega'_{x_i,A}\Omega_{x_i,A}\Delta^{-\frac{1}{2}})^{\frac{1}{2}}\mathbf{h}\rho||_2$, *where*

$$\mathbf{h} = HE\left[\left(R'_{K,H}(\mathcal{S}(l+\psi)(l+\psi)'\mathcal{S}'\otimes I_H)R_{K,H}\right)^{-1}R'_{K,H}\left(\mathcal{S}(l+\psi)\psi'\mathcal{S}^{-1}\otimes I_H\right)\right],$$
$$\rho = \left(\Delta^{-\frac{1}{2}}\otimes I_{K^2}\right)\mathrm{vec}\left(\mathbf{S}_{12}\right)/\sqrt{\Sigma_\delta}\ ,$$

$l = \mathbf{S}_2^{-\frac{1}{2}}\mathrm{vec}(C')$, $\psi = \mathbf{S}_2^{-\frac{1}{2}}(\eta_2 - \mathrm{vec}(C')) \sim \mathcal{N}(0, I_{KH})$, *and* $\mathcal{S} = ((\Delta/H)^{-\frac{1}{2}}\otimes I_H)\mathbf{S}_2^{\frac{1}{2}}$.

Proof. Since the expression in Proposition 1 is identical to that in Proposition 1 of Lewis and Mertens (2025a), after replacing that paper's K with this paper's H and that paper's N with this paper's K, the result follows immediately from Lemma 1 in that paper, after substituting in the value of the weighting matrix and scaling factor in the denominator of (34). □

Next, we define the concentration parameter, Λ:

Definition 2. *The concentration parameter for (30) is*

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

where $\Delta = R'_{K,H}(\mathbf{S}_2\otimes I_H)R_{K,H}$.

The concentration parameter is the (standardised) mean of the "denominator" of the 2SLS estimator and measures the signal-to-noise ratio in the first stage. Since it is the center of the distribution of the 2SLS denominator, it is critical for characterizing the weak instruments distribution of 2SLS estimators, and thus the asymptotic bias of $\hat{\beta}$, since it has the same limiting distribution. Note that, via Δ,

which sums variances over horizons, there is effectively a factor of H in the denominator (literally so in the case if $\mathbf{S}_2$ is homoskedastic across horizons), illustrating how adding horizons without identifying information weakens identification.

$\mathbb{B}_\tau(\mathbf{W})$ depends on the expectation of β^*. Stock and Yogo (2005) compute this expectation via numerical integration as a step towards computing the boundary of the weak instruments set. However, once $\mathbf{W}$ no longer has the Kronecker product form they assume under conditional homoskedasticity with serially uncorrelated errors, $\mathbf{W} = \Sigma \otimes I_H$, the value of this expectation depends on $\mathbf{W}$, so becomes regression-specific. To avoid conducting computationally extensive numerical integration for every regression estimated, we therefore exploit the Nagar approximation to $E[\beta^*]$, following Montiel Olea and Pflueger (2013) and Lewis and Mertens (2025a). This is the expectation of the second order Taylor approximation to β^*. Let B_n denote the Nagar approximation to B. The following proposition provides upper bounds on the Nagar bias that are independent of the nuisance paramters Λ and β, depending only on $\mathbf{W}$, which is consistently estimable.

Proposition 7. *Under Assumptions 1 and 2, the Nagar bias has the following bounds:*

(i) $B_n \leq B_n^*(\mathbf{W}, \lambda_{\min}) = \lambda_{\min}^{-1}\mathcal{B}(\mathbf{W})$, $\mathcal{B}(\mathbf{W}) = H^{-\frac{1}{2}}||(\Delta^{-\frac{1}{2}}\Omega'_{x_i,A}\Omega_{x_i,A}\Delta^{-\frac{1}{2}})^{\frac{1}{2}}||_2$
$\times \sup_{L_0 \in \mathbb{O}^{K\times H}} \{||M_1(I_N \otimes L_0 \otimes L_0)M_2\Gamma||_2\}$,

(ii) $B_n^*(\mathbf{W}, \lambda_{\min}) \leq \lambda_{\min}^{-1}\mathcal{B}^s(\mathbf{W})$, $\mathcal{B}^s(\mathbf{W}) = ||(\Delta^{-\frac{1}{2}}\Omega'_{x_j,A}\Omega_{x_j,A}\Delta^{-\frac{1}{2}})^{\frac{1}{2}}||_2 \min\{(2(K+1)/H)^{\frac{1}{2}}||M_2\Gamma||_2, ||\Gamma||_2\}$,

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

$$\Gamma = (\mathcal{S}\mathbf{W}_2^{-\frac{1}{2}}[\mathbf{W}_{12} : \mathbf{W}_2]' \otimes I_K)R_{K+1,H}(R'_{K+1,H}(\mathbf{W} \otimes I_H)R_{K+1,H})^{-\frac{1}{2}},$$
$$M_1 = R'_{K,K}\left(I_{K^3} + (\mathcal{H}_{K,K} \otimes I_K)\right), \; M_2 = R_{K,H}R'_{K,H}/(K+1) - I_{KH^2} \; .$$

Proof. Note that for a given β,

$$\Sigma_\delta = \sum_{h=0}^{H} \lim_{T\to\infty} \text{var}(\hat{\delta}_h) \tag{71}$$
$$= (1 \; -\beta') \left(\sum_{h=0}^{H-1} \text{avar}\left(\begin{bmatrix} \hat{\Psi}_h \\ \text{vec}(\hat{\Phi}_h) \end{bmatrix} \begin{bmatrix} \hat{\Psi}'_h & \text{vec}(\hat{\Phi}_h)' \end{bmatrix} \right) \right) (1 \; -\beta')' \tag{72}$$
$$= (1 \; -\beta')R'_{K+1,H}(\mathbf{W} \otimes I_H)R_{K+1,H}(1 \; -\beta')' \tag{73}$$

Note that this final expression is identical to the scale factor under the relative bias criterion in Lewis and Mertens (2025a), the first line of (A.8) in that paper. Therefore, the results follow from the proof of Theorem 1 in Lewis and Mertens (2025a), after adjusting notation and substituting in the value of the present weighting matrix. □

Part (i) of the theorem characterizes an upper bound on the Nagar bias, $B_n^*(\mathbf{W}, \lambda_{\min})$,

that only depends on $\mathbf{W}$ and $\lambda_{\min}$, the minimum eigenvalue of the concentration parameter in Definition 2. Importantly, $B_n^*(\mathbf{W}, \lambda_{\min})$ is often a sharp upper bound on the Nagar bias, by which we mean that it is attained for at least some value of β and eigenstructure of Λ. We call this upper bound $B_n^*(\mathbf{W}, \lambda_{\min})$ the "worst-case Nagar bias". The $\mathcal{B}(\mathbf{W})$ term in the expression for the worst-case Nagar bias is proportional to the largest singular value of a matrix that depends on the $KH-(K+1)/2$ nuisance parameters in the orthonormal matrix L_0, which needs to be obtained numerically. Part (ii) of the theorem provides an alternative simplified upper bound that requires no numerical optimization. This alternative bound is, in general, not sharp and (weakly) larger than the worst-case Nagar bias, but it is easy to obtain numerically even for very large-dimensional models – for instance, as H grows large, as in McKay and Wolf (2023).

These expressions for upper bounds on B_n depend only on a consistently estimable matrix ($\mathbf{W}$) and a scalar nuisance parameter $\lambda_{\min}$. Thus, our test takes the form

$$H_0 : \lambda_{\min} \leq \lambda_{\min}^*(\tau) \quad \text{vs.} \quad H_1 : \lambda_{\min} > \lambda_{\min}^*(\tau). \tag{74}$$

$\lambda_{\min}^*$ can be determined based on the bounds in Proposition 7. It thus remains to propose a test for whether $\lambda_{\min}$ exceeds that threshold. Proposition 8 provides such a test statistic and gives its limiting distribution.

Proposition 8. *Define the test statistic*

$$g_{\min} = \text{mineval}\{\hat{\Delta}^{-\frac{1}{2}}(m'\zeta_H\zeta_H'm)\hat{\Delta}^{-\frac{1}{2}}\},$$

where $\hat{\Delta} = R_{K,H}'(\hat{\mathbf{W}}_2 \otimes I_H)R_{K,H}$. *Then, under Assumptions 1 and 2,*

$$g_{\min} \xrightarrow{d} \text{mineval}\{R_{K,H}'(\Upsilon \otimes I_H)R_{K,H}/H\},$$

where the $KH \times KH$ *random matrix* $\Upsilon = \mathcal{S}(l+\psi)(l+\psi)'\mathcal{S}'$ *has a noncentral Wishart distribution,* $\Upsilon \sim \mathcal{W}(1, \Sigma_\Upsilon, Q)$, *with* 1 *degree of freedom, scale matrix* $\Sigma_\Upsilon = \mathcal{S}\mathcal{S}' \in \mathbb{P}^{KH}$, *and a matrix of noncentrality parameters* $Q = \Sigma_\Upsilon^{-1}\mathcal{S}ll'\mathcal{S}'$.[7]

Proof. Because the weak instruments limit of $\hat{\beta}$ is identical to that for the 2SLS estimator in Lewis and Mertens (2025a), the result is a restatement of Proposition 2 in that paper, after adjusting notation appropriately. □

Because this limiting distribution has an identical form to that in Lewis and Mertens (2025a), up to notation, it follows that a) a testing procedure using the supremum (over nuisance parameters) of critical values from that distribution is pointwise asymptotically valid (that paper's Proposition 3) b) while those critical values are infeasible, they can be well-approximated using the Imhof approximation

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

based on its first three cumulants c) those cumulants, which depend on nuisance parameters, can be weakly bounded above, with that bound depending only on $\lambda_{\min}$ and $\mathbf{W}$ (that paper's Theorem 2); d) an Imhof approximation based on those upper bounds will be conservative for the Imhof approximation based on the infeasible true cumulants (that paper's Proposition 4). In particular, the critical values can be constructed from the distribution

$$\mathcal{F}_{\mathcal{I}}(u;\kappa_1,\bar{\kappa}_2,\bar{\kappa}_3) = \mathcal{F}_{\chi^2_\varrho}((u-\kappa_1)4\omega+\varrho), \quad \varrho = 8\kappa_2\omega^2, \quad \omega = \kappa_2/\kappa_3,$$

where

$$\kappa_1 = H(1+\lambda_{\min} \tag{75}$$

$$\bar{\kappa}_n = 2^{n-1}(n-1)!\Big(\operatorname{maxeval}\{R'_{K,H}(\Sigma^n_\Upsilon \otimes I_H)R_{K,H}\} + nH\lambda_{\min}\operatorname{maxeval}\{\Sigma_\Upsilon\}^{n-1}\Big). \tag{76}$$