EVALUATING POLICY INSTITUTIONS

–150 YEARS OF US MONETARY POLICY–

REGIS BARNICHON
Federal Reserve Bank of San Francisco and CEPR

GEERT MESTERS
Federal Reserve Bank of New York

Abstract Given a loss function and a set of policy objectives, how should we evaluate the performance of policy institutions? In this work, we show that it is possible to evaluate policymakers with minimal assumptions about the underlying economic model. The distance to minimum loss—the component of the loss for which a policy institution can be held accountable—can (i) be computed from well-known and estimable sufficient statistics, namely the impulse responses to policy and non-policy shocks, and (ii) be decomposed into its shock-specific components. This decomposition reveals which macroeconomic shocks were most costly to overall performance and identifies directions for improvement. Applying the methodology to U.S. monetary policy since 1879, we find no material improvement over the first 100 years, with substantial and broad-based gains emerging only in the last 30 years.

KEYWORDS: optimal policy, reaction function, structural shocks, impulse responses, monetary history.

Regis Barnichon: Regis.Barnichon@sf.frb.org

Geert Mesters: geert.mesters@ny.frb.org

We thank our discussants David Romer and Frank Smets as well as Ryan Chahrour, Stephen Cecchetti, Davide Debortoli, Thomas Drechsel, John Friedman, Cosmin Ilut, Toru Kitagawa, Sylvain Leduc, Karel Mertens, Emi Nakamura, Franck Portier, Adam Shapiro, Jón Steinsson, Christian Wolf and numerous seminar participants for helpful comments. Mesters acknowledges financial support from the 2021 ERC Starting Grant POLICYMETRICS. The views expressed in this paper are the sole responsibility of the authors and do not necessarily reflect the views of the Federal Reserve Bank of San Francisco, the Federal Reserve Bank of New York or the Federal Reserve System.

1. INTRODUCTION

How should we evaluate and compare the performance of policy institutions? How should we evaluate and compare policy makers after their term in office? These questions are of central importance to the proper functioning of democratic and accountable institutions, yet there is little consensus on a suitable method for evaluating and comparing performance.

A naive approach would consist of measuring performance based on realized macroeconomic outcomes— that is, on the realized value of some loss function. For instance, we could assess a central banker based on average inflation and unemployment outcomes over her term. Unfortunately, that approach suffers from numerous confounding problems, as many features of the economy —the economic environment— are outside policy makers' control yet affect performance: (i) different policy makers may face different initial conditions upon beginning their term (e.g., a central banker may inherit a strong or weak economy from her predecessor), (ii) different policy makers may face different economic disturbances (e.g., a central banker may experience a financial crisis or an energy price shock that will affect her ability to stabilize inflation and unemployment), and (iii) different policy makers may live in different economies (e.g., a steeper or flatter Phillips curve will affect a central banker's ability to control inflation). In sum, the economic environment can confound performance.

So, what makes a policy institution good or bad? Our starting point is simple: policy makers react to the state of the world by taking actions to minimize a loss function, i.e. they use their policy instruments to achieve certain policy objectives. A policy maker's reaction to the state of the world can be expressed as a policy rule —a reaction function—, and a policy maker is best performing, when her reaction function is "optimal", i.e., delivers the minimum loss possible given the environment. In that context, evaluating a policy maker requires measuring the distance between the policy maker's rule and the optimal rule, that is measuring the distance to the minimum loss possible *given* the environment.

To characterize the minimum feasible loss in a given economic environment, one approach is to use a structural model fitted to data spanning the policy maker's term, derive the optimal rule and compute the associated minimum loss. A risk with this approach, however, is model misspecification: if the model is misspecified, evaluations may be inaccurate.

In this paper, we propose a semi-structural method to evaluate policy makers with minimal assumptions about the underlying economic model. Specifically, for a large class of linear forward looking macro models and quadratic loss functions, it is possible to measure the distance to the optimal reaction function —and distance to minimum loss— from well known and estimable sufficient statistics: the impulse responses to policy and non-policy shocks.

To characterize the optimal policy rule without having to rely on a specific underlying model, we exploit two new results. First, an identification result: knowledge of the optimal reaction to structural shocks alone is sufficient to characterize the optimal policy rule—that is, to construct a policy rule that minimizes the loss function given the environment. Second, a sufficient statistics result: the optimal reaction to structural shocks can be characterized from the impulse responses to policy and non-policy shocks.

Taken together, these results imply that the impulse responses to policy and non-policy shocks are sufficient to construct policy evaluation statistics at the shock level: the distance to the optimal policy response to a particular structural shock (say a financial shock, an oil shock, etc.) and the distance to minimum loss for that particular shock. The total distance to minimum loss —a measure of overall performance— can then be obtained by simply summing the shock-specific distances to minimum loss.

An attractive feature of our constructive formulation of the total distance to minimum loss is that it enables researchers to not only evaluate overall performance but also to understand the reasons behind suboptimal performances. By decomposing the total DML into its shock-specific component, researchers can isolate which types of shocks were most costly to overall performance and to identify directions for improvement.

In a dynamic setting, computing the total distance to minimum loss requires identifying all policy and non-policy news shocks across all possible horizons. Since this data requirement is rarely met in practice, we show how subset statistics, which rely on only a subset of shocks, can be used to evaluate policymakers, subject to two qualifications.

First, when a researcher cannot identify all shocks affecting a policymaker during her term, the overall evaluation may not be exhaustive, as some shocks are missing. To address this "missing shocks" problem, we propose two solutions: (i) we derive bounds on the total distance to minimum loss (using the same set of sufficient statistics and without additional assumptions), or (ii) we show how it is possible directly estimate the total DML from oracle forecast innovations instead of structural shocks. The latter approach, however, requires an additional assumption —the ability to construct oracle forecasts.

Second, when a researcher cannot identify all types of policy shocks (for instance, forward guidance shocks at all different horizons), the characterization of the optimal reaction function is only "partial", as some possible policy path responses are not considered. To give a concrete example, say we only identify policy shocks that affect the short-end of the policy path, then we can only explore the optimality of the short end of the policy path response to shocks. This does not invalidate the policy evaluation, but this is a limitation to keep in mind when interpreting the results.

We apply our methodology to study the performance of US monetary policy over the past 150 years and revisit many important questions regarding the conduct of monetary policy: (i) Did monetary policy improve since the time of the Great Depression? Is the Great Moderation post-Volcker a sign of good policy or simply the outcome of good luck? (e.g., Clarida et al., 2000, Galı et al., 2003, Galí and Gambetti, 2009)? (ii) While many observers agree that monetary policy was superior during the 2007-2009 financial crisis than during the 1929-1933 financial crisis (e.g., Wheelock et al., 2010, Almunia et al., 2010), can we confirm and quantify this improvement? In other words, did Bernanke fulfill his promise to Milton Friedman when he said that the Fed "won't do it again", i.e., won't repeat the mistakes of the Great Depression (Bernanke, 2002)? (iii) How does the post World War II Fed and the de-anchoring of inflation expectations compare to de-anchoring of inflation expectations in the interwar Fed period (Romer and Romer, 2013)? (iv) Finally, did the founding of the Federal Reserve in 1913 lead to better macroeconomic outcomes compared to the passive gold standard era (e.g., Bordo and Kydland, 1995)?

To assess monetary policy performance across historical periods, we construct and decompose the total DML into the separate contributions of six macro shocks: (i) financial shocks, (ii) government spending shocks, (iii) energy price shocks, (iv) inflation expectation shocks, (v) productivity shocks and (iv) monetary shocks, all identified using the state-of-the-art in the empirical macro literature. We consider US monetary policy over four distinct periods: (a) 1879-1912 covering the Gold standard period until the founding of the Federal Reserve, (b) 1913-1941 covering the early Fed years to the US entering World War II, (c) 1954-1984 covering the post World War II period until the beginning of the Great Moderation, and (d) 1990-2019 covering the Great Moderation period, the financial crisis and up to the COVID crisis. For the Gold Standard period, the identification of monetary shocks is less developed, and we propose a new identification strategy based on large gold mine discoveries.

Given a loss function that places equal weight on inflation and unemployment, our main results are as follows: (i) we estimate large and uniform improvements in the conduct of monetary policy, but *only* in the last 30 years, (ii) we cannot reject that the Fed's reaction to recent financial shocks (notably the 2007-2008 financial crisis) was appropriate, in contrast to the "highly" sub-optimal reaction of the Fed to the financial shocks of the Great Depression, (iii) the Fed's reaction function during the 1960s-1970s is almost as sub-optimal as the reaction function of the early Fed, though the nature of the shocks is different, and (iv) the founding of the Fed initially led to worse performance than the passive Gold Standard. In particular, faced with financial or government spending shocks, the Fed fared worse than the passive Gold Standard.

Related literature An early contribution is Fair (1978) who highlights the distortions stemming from different initial conditions and economic environments. His solution was to adopt optimal control methods to compare policymakers through the lens of a fully specified model. Modern versions of this approach include Galı et al. (2003), Gali and Gertler (2007), Blanchard and Galí (2007). Unfortunately, specifying the correct model for the policy rule or the non-policy block is a difficult task (e.g., Svensson, 2003, Mishkin, 2010). A less structural approach has studied monetary performance through the lens of estimated policy rules —requiring only the specification of a policy rule— (e.g. Taylor, 1999, Clarida et al., 2000, Orphanides, 2003, Coibion and Gorodnichenko, 2011). In particular, a number of studies compared the Fed in the pre- and post-Volcker periods by assessing whether the Taylor principle was satisfied. However, beyond the Taylor principle, that approach can say little about the optimality of reaction functions, and thus can only provide a coarse evaluation of reaction functions.

In the context of fiscal policy, Blinder and Watson (2016) improve on the naive approach of policy evaluation —measuring performance based on unconditional realized outcomes— by *projecting out* specific shocks, i.e., by trying to control for good (or bad) luck. In contrast, our approach *projects on* the space spanned by specific non-policy shocks and studies performance in that space.

Closer to our work, the literature has proposed reduced-form methods to study policy rule counterfactuals (e.g., Sims and Zha, 2006, Bernanke et al., 1997, Leeper and Zha, 2003), though these approaches are not fully robust to the Lucas critique. Instead, our approach builds on recent work showing that robustness to the Lucas critique is possible in a large class of macroeconomic models (McKay and Wolf, 2023, Caravello et al., 2024): When the coefficients of the non-policy block are independent of the coefficients of the policy block, it is possible to reproduce any policy rule counterfactual with an appropriate combination of policy news shocks at different horizons. Our work exploits a little-studied class of policy rule counterfactuals —counterfactual reactions to non-policy shocks—, which have appealing properties: (i) the class is sufficient to characterize the optimal reaction function, (ii) the class allows to split the optimal policy problem into computationally simple separate problems, allowing us to evaluate policymakers under subset identification, and (iii) each sub-problem has an economic interpretation; allowing us to understand the sources of sub-optimal policy decisions, for instance the types of shocks that policymakers could have handled better. This last property allows us to relate to and quantify a large literature on previous policy misses, notably the seminal narrative studies of US monetary policy (Friedman and Schwartz, 1963, Meltzer, 2009).

Lastly, our paper relates to the sufficient statistics approach for macro policy proposed in Barnichon and Mesters (2023). Different from our focus on reaction function evaluation, Barnichon and Mesters (2023) focus on the time t optimal policy problem —how to set the policy path today given the state of the economy—, instead of the unconditional policy problem that we consider here —how to set up the policy rule to minimize the unconditional loss—. Barnichon and Mesters (2023) show that the characterization of the optimal targeting policy rule can be reduced to the estimation of two sufficient statistics (i) the impulse responses of the policy objectives to policy shocks, and (ii) oracle forecasts for the policy objectives conditional on some baseline policy rule. This paper uses a different set of sufficient statistics—policy *and* non-policy shocks—, providing an economic interpretation for the sources of optimization failures, as discussed above.

2. ILLUSTRATIVE EXAMPLE

Before formally describing our general framework, we first illustrate how it is possible to evaluate and compare policy makers without having access to the underlying economic model nor the policy rule. To illustrate the method, we take a baseline New Keynesian (NK) model, which allows to highlight the main mechanisms of our approach.

The log-linearized Phillips curve and intertemporal (IS) curve of the baseline New-Keynesian model are given by

$$\pi_t = \mathbb{E}_t \pi_{t+1} + \kappa x_t + \sigma_\xi \xi_t \ , \tag{1}$$

$$x_t = \mathbb{E}_t x_{t+1} - \frac{1}{\sigma}(i_t - \mathrm{E}_t \pi_{t+1}) \ , \tag{2}$$

with π_t the inflation gap, x_t the output gap, i_t the nominal interest rate set by the central bank and $\sigma_\xi \xi_t$ an iid cost-push shock with mean zero and variance σ_ξ^2. The parameters are collected in $\theta = (\kappa, \sigma, \sigma_\xi)'$. We can think of θ as capturing the economic "environment"; the slopes of the (PC) and (IS) curves and the standard deviation of the non-policy shock (here, the cost-push shock).

The policy maker sets the interest rate following the rule

$$i_t = \phi_\pi \pi_t + \sigma_\epsilon \epsilon_t \ , \tag{3}$$

where $\sigma_\epsilon \epsilon_t$ is an iid policy shock with mean zero and variance σ_ϵ^2, and $\phi = (\phi_\pi, \sigma_\epsilon)$ is a vector of policy parameters, or the reaction function. For $\phi_\pi > 1$ we can solve the model and express the endogenous variables $Y_t = (\pi_t, x_t)'$ and i_t as functions of the exogenous shocks $S_t = (\xi_t, \epsilon_t)'$, i.e.

$$Y_t = \Theta S_t \qquad \text{and} \qquad i_t = \Theta_p S_t \ , \tag{4}$$

where $\Theta = (\Gamma, \mathcal{R})$ and $\Theta_p = (\Gamma_p, \mathcal{R}_p)$ capture the impulse responses of the non-policy variables (Y) and the policy variable (i) to the cost-push shock (Γ, Γ_p) and the monetary policy shock ($\mathcal{R}, \mathcal{R}_p$). We emphasize that each impulse response depends on the environment θ and the reaction function ϕ.[1]

[1]Formally, we have

$$\mathcal{R} = \sigma_\varepsilon \begin{bmatrix} \frac{-\kappa/\sigma}{1+\kappa\phi_\pi/\sigma} \\ \frac{-1/\sigma}{1+\kappa\phi_\pi/\sigma} \end{bmatrix} \ , \quad \Gamma = \sigma_\xi \begin{bmatrix} \frac{1}{1+\kappa\phi_\pi/\sigma} \\ \frac{-\phi_\pi/\sigma}{1+\kappa\phi_\pi/\sigma} \end{bmatrix} \ , \quad \mathcal{R}_p = \sigma_\varepsilon \frac{1}{1+\kappa\phi_\pi/\sigma} \ , \quad \Gamma_p = \sigma_\xi \frac{\phi_\pi}{1+\kappa\phi_\pi/\sigma} \ ,$$

Evaluating policy makers requires taking a stance on a performance metric. To that effect, we consider the loss function

$$\mathcal{L} = \mathbb{E}Y_t'Y_t \,, \qquad \text{which using (4) becomes } \mathcal{L} = \mathrm{Tr}(\Theta'\Theta) = \Gamma'\Gamma + \mathcal{R}'\mathcal{R} \,, \tag{5}$$

where $\mathrm{Tr}()$ denotes the trace operator. An "optimal reaction function" is then defined as any ϕ that minimizes $\mathcal{L}(\phi;\theta)$ given the underlying structure of the economy, i.e., given equations (1)-(3), and expression (5) shows that minimizing the unconditional loss $\mathcal{L}$ is equivalent to minimizing the sum-of-squares of the impulse responses of shocks hitting the economy, here Γ and $\mathcal{R}$. An optimal policy is thus a policy rule that best mutes the effects of shocks on average.

In this example the optimal reaction function is unique and given by $\phi^{\mathrm{opt}} = (\phi_\pi^{\mathrm{opt}}, \sigma_\epsilon^{\mathrm{opt}})' = (\kappa\sigma, 0)'$ (e.g. Galí, 2015). First, exogenous policy changes are not optimal, and an optimal policy features no policy shocks ($\sigma_\epsilon^{\mathrm{opt}} = 0$). Second, the optimal reaction coefficient ϕ_π^{opt} is the coefficient that minimizes the effects of cost-push shocks, i.e., that best mutes Γ. The minimum loss is then given by

$$\mathcal{L}^{\mathrm{opt}} = \Gamma^{\mathrm{opt}'}\Gamma^{\mathrm{opt}}, \tag{6}$$

with Γ^{opt} being equal to Γ evaluated at $(\phi^{\mathrm{opt}}, \theta)$, i.e. the minimal effect of cost-push shocks that a policy maker can achieve given the environment θ.

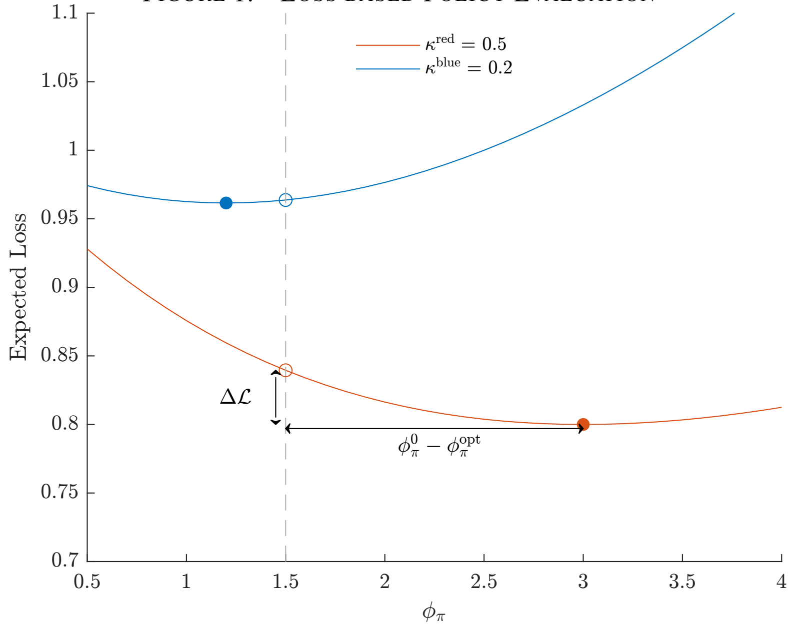
A naive approach to policy evaluation

Consider a policy maker with reaction function ϕ^0 during her term and associated loss $\mathcal{L}^0$. How should we evaluate that policy maker?

A naive approach would consist in comparing realized losses. Specifically, (i) compute the average loss during a policy maker's term, which provides an estimate of the loss $\mathcal{L}^0$, and (ii) evaluate and rank policy makers based on that estimate. Policy makers with higher average loss would then be deemed less performant. Unfortunately, the parameter vector θ that describes the economic environment acts as a confounder by influencing the impulse responses to shocks and thus the loss, see (4)-(5).

To illustrate how the economic environment can distort such naive policy evaluation, consider two policy makers —Red and Blue— following the same rule $\phi_\pi^0 = 1.5$ but operating in different environments: one with a steeper Phillips curve ($\kappa^{\mathrm{red}} = 0.5$) and the other

FIGURE 1.—LOSS BASED POLICY EVALUATION



Note: The panel depicts the original loss function $\mathcal{L}(\phi;\theta)$ with $\phi = (\phi_\pi, 0)$ and traces the loss as a function of ϕ_π, fixing $\sigma = 6$, $\sigma_\epsilon = 0$ and $\sigma_\xi = 1$. The optimal rule ϕ_π^{opt} is indicated by the filled dots. The empty dots indicated the policy maker's policy rule ϕ_π^0.

with a flatter Phillips curve ($\kappa^{\mathrm{blue}} = 0.2$). Figure 1 plots the loss function as a function of the rule parameter ϕ_π for these two policy makers. The empty dot marks their actual policy rule (here $\phi_\pi^0 = 1.5$ for both Red and Blue).

Since the loss for Red is lower than under Blue, a naive approach to policy evaluation would conclude that Red is a better policy maker than Blue. However, it's the exact opposite: in this example, Red is further away from the optimal reaction function than Blue. The filled dots mark the optimal reaction function ϕ^{opt} for each policy maker, and the distance to the optimal reaction function $\phi^0 - \phi^{\mathrm{opt}}$ —the horizontal distance between the filled dot and the empty dot—, is larger for Red than for Blue. In other words, Red performs less well than Blue. The reason for these different conclusions is the underlying environment: in the steep Phillips curve world of Red, it is *easier* to achieve a lower loss.

To properly compare Red and Blue, we must thus take into account the environment, i.e., measure the distance to the minimum feasible loss given the environment; the distance $\Delta\mathcal{L}$ in Figure 1. To do so, one possible approach consists in specifying a structural model, fit that model to the data spanning the policy maker's term and then compute the optimal reaction coefficient ϕ^{opt} and the associated minimum feasible loss $\mathcal{L}^{\text{opt}}$ from that model. In this example this amounts to specifying the Phillips and IS curves and estimating the associated parameters θ. A risk with this approach however is model mis-specification: if the model does not capture the full complexity of the underlying environment, the policy assessment can be compromised. In this paper, we propose a different approach, an approach that requires minimal assumptions on the underlying economic model.

A sufficient statistics approach to policy evaluation

A class of policy rule counter-factuals To outline our approach we start from a simple idea: instead of minimizing the loss with respect to the reaction coefficients in front of endogenous variables (here ϕ_π) as is common in the literature (Galí, 2015), we propose to optimize with respect to the reaction coefficients in front of structural shocks. While this class of rule counterfactuals could seem of little direct interest, they have two important, yet overlooked, properties: (i) the effects of counterfactual reaction to structural shocks can be computed with minimal assumptions on the underlying model, depending only on estimable sufficient statistics, and (ii) the optimal reaction to structural shocks is sufficient to fully characterize the optimal policy rule and to compute the minimum attainable loss.

To see that, denote by $\phi^0 = \left(\phi^0_\pi, \sigma^0_\epsilon\right)$ the policy maker's reaction function and consider the policy rule counter-factual

$$i_t = \phi^0_\pi \pi_t + \underbrace{\sigma^0_\epsilon(\tau_\xi \xi_t + \tau_\epsilon \epsilon_t)}_{\text{Reaction adjustment}} + \sigma^0_\epsilon \epsilon_t \ , \tag{7}$$

where $\tau = (\tau_\xi, \tau_\epsilon)$ is a vector of responses to structural shocks. Unlike the original reaction function (3), the modified reaction function (7) fixes the reaction coefficients ϕ^0 at their baseline value.

Following the same steps that led to (4), we can solve the model under that modified policy rule and express the endogenous variables as a function of exogenous shocks to get

$$Y_t = (\Theta^0 + \mathcal{R}^0\tau)S_t \qquad \text{and} \qquad i_t = (\Theta^0_p + \mathcal{R}^0_p\tau)S_t \tag{8}$$

where the ^{0} superscript indicates that the impulse response is computed under (ϕ^0, θ), such that $\Theta^0 = (\Gamma^0, \mathcal{R}^0)$. From expression (8), we can see how the rule adjustment τ modifies the impulse response to shocks. Take for instance Γ^0, the effect of the cost-push shock. A reaction adjustment τ_ξ changes the effect from Γ^0 to $\Gamma^0 + \mathcal{R}^0\tau_\xi$, which means that the "old" impulse responses Γ^0 and $\mathcal{R}^0$ are all we need to compute the effects of the policy rule counter-factual (7). A similar result holds for how τ_ϵ modifies the impulse response $\mathcal{R}^0$. Intuitively, a change τ in the reaction to an exogenous variable (here xi or ϵ) has equilibrium effects that can be measured by the causal effects of policy shocks, i.e. $\mathcal{R}^0$.

Optimal reaction adjustment From (8), we can use Γ^0 and $\mathcal{R}^0$ to search for the optimal reaction coefficient to structural shocks. To that effect, consider an auxiliary loss function that takes τ as its argument while holding ϕ^0 fixed:

$$\begin{aligned}\mathsf{L}(\tau) &= \mathbb{E}(Y_t'Y_t) \qquad \text{with} \quad Y_t = (\Theta^0 + \mathcal{R}^0\tau)S_t \\ &= \mathrm{Tr}[(\Theta^0 + \mathcal{R}^0\tau)'(\Theta^0 + \mathcal{R}^0\tau)]\end{aligned}$$

Solving for the optimum reaction adjustment $\tau^* = (\tau_\xi^*, \tau_\epsilon^*) = \arg\min_\tau \mathsf{L}(\tau)$, we get[2]

$$\tau_\xi^* = -(\mathcal{R}^{0'}\mathcal{R}^0)^{-1}\mathcal{R}^{0'}\Gamma^0 \quad \text{and} \quad \tau_\epsilon^* = -1 \ . \tag{9}$$

and

$$\mathsf{L}(\tau^*) = \mathcal{L}^{\text{opt}} \ . \tag{10}$$

In other words, optimizing with respect to τ is sufficient to fully characterize the optimal reaction function, and the auxiliary loss function has the same minimum as the original loss function $\mathcal{L}$. The statistic τ^* is the optimal reaction adjustment:[3] (i) τ_ξ^* modifies the policy rule coefficient for non-policy shocks ξ_t in order to reach Γ^{opt}; the minimal effect

[2]To show $\mathsf{L}(\tau^*) = \mathcal{L}^{\text{opt}}$, plug in τ^* into the auxiliary loss function to obtain

$$\begin{aligned}\mathsf{L}(\tau^*) &= \Gamma^{0'}(I - \mathcal{R}^0(\mathcal{R}^{0'}\mathcal{R}^0)^{-1}\mathcal{R}^{0'})\Gamma^0 \\ &= \frac{\sigma_\xi^2}{1+\kappa^2} = \mathcal{L}^{\text{opt}} \ ,\end{aligned}$$

using the expressions for Γ^0 and $\mathcal{R}^0$ defined in footnote 1.

[3]Note how τ_ξ^* is the coefficient of a regression of Γ^0 on $-\mathcal{R}^0$; a regression in impulse response space. Intuitively, Γ^0 (the impulse response to a cost-push shock) captures what the policy maker *did* on average to counteract cost-push shocks with his rule ϕ^0, while $\mathcal{R}^0$ (the impulse response to a monetary shock) captures what the policy maker *could have done* to counteract cost-push shocks —how reacting to ξ_t by τ could have better stabilized the

of cost-push shock that a policy maker can achieve given the environment θ,[4] and (ii) τ_ϵ^* cancels monetary mistakes by setting policy shocks back to zero with $\tau_\epsilon^* = -1$.

While the optimal reaction adjustment allows to fully characterize the optimal reaction function, it does not have a transparent economic interpretation. Instead, we can compute the corresponding optimal policy response adjustment $\Delta\Theta_p = \Theta_p^0 - \Theta_p^{\text{opt}}$, which captures how the systematic policy responses to shocks should be adjusted. Using (8) we have that

$$\Delta\Theta_p = -\mathcal{R}_p^0 \tau^* \ , \tag{11}$$

as $\Theta_p^{\text{opt}} = \Theta_p^0 + \mathcal{R}_p^0 \tau^*$. The columns of $\Delta\Theta_p$ capture the optimal correction to the policy response to each type of shock (ξ or ϵ), and they allow to assess the "quality" of the policy maker's response to each type of shock, with larger corrections indicating a poorer reaction function.

Distance to minimum loss With the optimal reaction adjustment τ^* in hand, it is then straightforward to get expressions for (i) $\Delta\mathcal{L}$ the distance to minimum loss, and (ii) decompose $\Delta\mathcal{L}$ into the contribution of each structural shock. We have

$$\Delta\mathcal{L} = \mathcal{L}^0 - \mathcal{L}^{\text{opt}} = \Delta\mathcal{L}_\xi + \Delta\mathcal{L}_\epsilon \ , \tag{12}$$

where

$$\Delta\mathcal{L}_\xi = \Gamma^{0\prime}\mathcal{R}^0 \left(\mathcal{R}^{0\prime}\mathcal{R}^0\right)^{-1} \mathcal{R}^{0\prime}\Gamma^0 \qquad \text{and} \qquad \Delta\mathcal{L}_\epsilon = \mathcal{R}^{0\prime}\mathcal{R}^0 \ . \tag{13}$$

Intuitively, each optimal reaction adjustment focuses on a different structural shock, so that each adjustment assesses a separate dimension of policy performance.

Decomposition (12) is at the core of our approach to policy evaluation. From (12)-(13), we can construct the *total* distance to minimum loss $\Delta\mathcal{L}$ and assess *overall* performance, while expression (12) allows to decompose that total distance into separate shock-specific components, revealing which shocks were most costly to overall performance.

---

effect of cost-push shocks by transforming Γ^0 into $\Gamma^0 + \tau\mathcal{R}^0$, see (8)—. A regression on $\mathcal{R}^0$ on Γ^0 precisely finds the τ that minimizes the sum-of-squares of that adjusted impulse response, i.e., that best cancels out the effects of non-policy shock. At an optimal policy rule, Γ^0 and $\mathcal{R}^0$ should be orthogonal.

[4] In this example, we have $\tau_\xi^* = \frac{\sigma_\xi}{\sigma_\epsilon} \frac{-\kappa\sigma + \phi_\pi}{(1+\kappa^2)^2}$.

On the duality of policy performance evaluation Note how we derived two complementary statistics to measure performance: (i) the optimal policy response adjustment ($\Delta\Theta_p$), and (ii) the distance to minimum loss ($\Delta\mathcal{L}$). The two statistics play complementary roles, each capturing a different side of the same "performance coin". The optimal policy response adjustment measures the extent of good policy —it directly assesses the reaction function—, while the distance to minimum loss ($\Delta\mathcal{L}$) captures what we ultimately care about —the "welfare" consequences of good/bad policy. Unlike $\Delta\Theta_p$ however, the distance to minimum loss is not fully under control of the policy maker, as the same gap $\Delta\Theta_p$ could imply larger or smaller welfare losses depending on the underlying structural parameters and shocks variance. This is the luck aspect of policy performance.

In sum, this example illustrates how we can evaluate and compare policy makers without specifying an explicit reaction function nor a specific structural macro model. Instead, the only requirement is to estimate two sufficient statistics: the impulse responses Γ and $\mathcal{R}$ over a policy maker's term. The next sections show that these findings continue to hold for general linear forward looking macro models.

3. ENVIRONMENT

We describe a general stationary macro environment for a policy maker (or institution) who faces an infinite horizon economy. To describe the economy we distinguish between two types of observable variables: non-policy variables $y_t \in \mathbb{R}^{M_y}$ and the policy instrument $p_t \in \mathbb{R}$. The policy instrument is different from the other variables as it is under the direct control of the policy maker. To describe the economy we use a sequence space representation (Auclert et al., 2021). Let $\mathbf{Y} = (y_0', y_1', \ldots)'$ denote the paths for the non-policy variables and $\mathbf{P} = (p_0, p_1, \ldots)'$ denote the path for the policy variable p. For instance, one can think of $\mathbf{P}$ as the path of the policy rate for a central bank, while $\mathbf{Y}$ would typically comprise the paths of the inflation and unemployment gaps.[5]

[5] Generalizing our approach to multiple policy instruments ($p_t \in \mathbb{R}^{M_p}$) is a straightforward extension, but focusing on only one policy instrument clarifies the exposition, as the policy choice reduces to setting a vector: the policy path.

Working under perfect foresight, we consider a generic model for the paths of the endogenous variables

$$\begin{aligned} \mathcal{A}_{yy}\mathbf{Y} - \mathcal{A}_{yp}\mathbf{P} &= \mathcal{B}_{y\xi}\boldsymbol{\Xi} \\ \mathcal{A}_{pp}\mathbf{P} - \mathcal{A}_{py}\mathbf{Y} &= \mathcal{B}_{p\xi}\boldsymbol{\Xi} + \mathcal{B}_{p\epsilon}\boldsymbol{\epsilon} \end{aligned} , \tag{14}$$

where $\boldsymbol{\epsilon} = (\epsilon_0, \epsilon_1, \dots)'$ and $\boldsymbol{\Xi} = (\xi_0', \xi_1', \cdots)'$ are sequences of policy and non-policy shocks, respectively. The first equation captures the non-policy block of the economy, while the second equation captures the policy rule.

We normalize all elements of Ξ and $\boldsymbol{\epsilon}$ to have mean zero and unit variance. Also, we assume that they are serially and mutually uncorrelated, consistent with the common definition of structural shocks (e.g. Bernanke, 1986, Ramey, 2016). The structural maps $\mathcal{A}_{\cdot\cdot}$ and $\mathcal{B}_{\cdot\cdot}$ are conformable and may depend on underlying structural parameters. We split them in two parts: the economic environment $\theta = \{\mathcal{A}_{yy}, \mathcal{A}_{yp}, \mathcal{B}_{y\xi}\}$ which the policy maker takes as given, and the reaction function $\phi = \{\mathcal{A}_{pp}, \mathcal{A}_{py}, \mathcal{B}_{p\xi}, \mathcal{B}_{p\epsilon}\}$, which is under the control of the policy maker and we assume that $\mathcal{B}_{p\epsilon}$ is invertible. Further, we impose that ϕ and θ are independent in the sense that $\partial\theta_i/\partial\phi_j = 0$ for all entries i, j, i.e. changing the reaction function does not directly change the coefficients θ and all effects of ϕ on $\mathbf{Y}$ go via the policy path $\mathbf{P}$.

We denote by Φ the set of all reaction functions ϕ for which the model (14) implies a unique equilibrium, that is all ϕ for which

$$\mathcal{A} = \begin{pmatrix} \mathcal{A}_{yy} & \mathcal{A}_{yp} \\ \mathcal{A}_{py} & \mathcal{A}_{pp} \end{pmatrix} \qquad \text{is invertible.}$$

Many structural models found in the literature can be written in the form of (14); prominent examples include New Keynesian models and heterogeneous agents models.

For any $\phi \in \Phi$ we can write the expected path of the non-policy variables as a linear function of the policy and non-policy shocks.

LEMMA 1: *Given the generic model* (14) *with* $\phi \in \Phi$*, we have*

$$\mathbf{Y} = \Theta(\phi, \theta)\mathbf{S} \qquad \textit{and} \qquad \mathbf{P} = \Theta_p(\phi, \theta)\mathbf{S} , \tag{15}$$

where $\mathbf{S} = (\boldsymbol{\Xi}', \boldsymbol{\epsilon}')'$ *with conforming partition of the impulse response maps* $\Theta(\phi, \theta) = (\Gamma(\phi, \theta), \mathcal{R}(\phi, \theta))$ *and* $\Theta_p(\phi, \theta) = (\Gamma_p(\phi, \theta), \mathcal{R}_p(\phi, \theta))$.

Explicit characterizations for the impulse response maps are given in the appendix. Note the similarity between (15) and (4), as the illustrative static NK model is a special case with only contemporaneous shocks. Lemma 1 implies that the identification of the impulse responses requires observing part of the *future* shocks in Ξ and $\boldsymbol{\epsilon}$. That is, news shocks at the different horizons are needed for identification. The supplementary material spells out this point out in more detail.

Evaluation criteria

We consider a researcher who is interested in evaluating a policymaker based on their success at stabilizing some subset of the non-policy variables y_t around certain desired targets y^* for some time periods $t = 0, 1, 2, \ldots$. For ease of notation we set the targets to zero, as we can think of y_t as being defined as deviations from the desired targets.

We measure performance using the loss function

$$\mathcal{L}(\phi; \theta) = \mathbb{E}\mathbf{Y}'\mathcal{W}\mathbf{Y} \ , \tag{16}$$

where $\mathcal{W}$ is a positive semi-definite weighting matrix, which selects and weights the specific variables and horizons that are part of the researcher's evaluation criteria. The loss (16) is the researcher's evaluation criterion for scoring policymaker performance —an input into our framework—.

Using Lemma 1, we can rewrite the loss function as

$$\mathcal{L}(\phi; \theta) = \mathrm{Tr}(\Theta(\phi, \theta)'\mathcal{W}\Theta(\phi, \theta)) \ . \tag{17}$$

As in the simple example, minimizing the unconditional loss $\mathcal{L}$ is thus equivalent to minimizing the (weighted) sum-of-squares of the impulse responses to the different shocks affecting the economy. Intuitively, minimizing this loss can be seen as a *timeless* definition of optimal policy —representing the loss of a policymaker appointed at the beginning of time and in place forever—, and an optimal policy is a policy rule that best dampens the effects of shocks. This interpretation will be useful in getting some intuition for our sufficient statistics formula.

The actions of the policymaker are summarized by the reaction function ϕ. We define a reaction function to be optimal if it minimizes the loss function (16). Formally, the set of optimal reaction functions is given by

$$\Phi^{\text{opt}} = \left\{ \phi : \phi \in \underset{\phi \in \Phi}{\operatorname{argmin}} \, \mathcal{L}(\phi;\theta) \quad \text{s.t.} \quad (14) \right\} . \tag{18}$$

The definition implies that we only consider optimal reaction functions that lie in Φ; the set of reaction functions that imply a unique equilibrium. We denote by ϕ^{opt} an arbitrary reaction function in Φ^{opt}.

4. POLICY EVALUATION WITH SUFFICIENT STATISTICS

In this section, we show how we can evaluate a policymaker with a reaction function ϕ^0 by measuring (i) the distance to minimum loss ($\Delta\mathcal{L}$) and (ii) the optimal adjustment of the policy path response to shocks ($\Delta\Theta_p$):

$$\Delta\mathcal{L} = \mathcal{L}^0 - \mathcal{L}^{\text{opt}} \qquad \text{and} \qquad \Delta\Theta_p = \Theta_p^0 - \Theta_p^{\text{opt}} , \tag{19}$$

where $\Theta_p^0 = \Theta_p(\phi^0, \theta)$ and $\mathcal{L}^0 = \mathcal{L}(\phi^0;\theta)$ are evaluated under the policymaker's reaction function and $\mathcal{L}^{\text{opt}} = \mathcal{L}(\phi^{\text{opt}};\theta)$ and $\Theta_p^{\text{opt}} = \Theta_p(\phi^{\text{opt}}, \theta)$ are evaluated under an optimal reaction function.

4.1. *Optimal reaction adjustment and distance to minimum loss*

Following the same steps as the simple example of Section 2, we propose to characterize the optimal rule by considering a thought experiment in which we adjust the policymaker's reaction coefficients to the structural shocks. Specifically, consider the auxiliary reaction function

$$\mathcal{A}_{pp}^0 \mathbf{P} - \mathcal{A}_{py}^0 \mathbf{Y} = \mathcal{B}_{p\xi}^0 \boldsymbol{\Xi} + \mathcal{B}_{p\epsilon}^0 \boldsymbol{\epsilon} + \mathcal{B}_{p\epsilon}^0 \mathcal{T} \mathbf{S} , \tag{20}$$

where $\mathcal{T}$ is a map of reaction adjustments to the shocks $\mathbf{S}$. Each element of $\mathcal{T}$ corresponds to a different counterfactual rule, whereby we modify how different horizons of the policy path respond to one specific macro shock. The following lemma establishes how a rule adjustment $\mathcal{T}$ affects the equilibrium allocation.

LEMMA 2: *Consider the generic model* (14) *with* $\phi^0 \in \Phi$ *and the modified policy rule* (20). *We have*

$$\mathbf{Y} = (\Theta^0 + \mathcal{R}^0\mathcal{T})\mathbf{S} \qquad \textit{and} \qquad \mathbf{P} = (\Theta_p^0 + \mathcal{R}_p^0\mathcal{T})\mathbf{S} \,, \tag{21}$$

where $\Theta^0 \equiv \Theta(\phi^0, \theta)$, $\Theta_p^0 = \Theta_p(\phi^0, \theta)$, $\mathcal{R}^0 \equiv \mathcal{R}(\phi^0, \theta)$ *and* $\mathcal{R}_p^0 \equiv \mathcal{R}_p(\phi^0, \theta)$.

The reaction adjustment $\mathcal{T} = (\mathcal{T}_\xi, \mathcal{T}_\epsilon)$ affects the equilibrium by changing the impulse responses to non-policy shocks from Γ^0 to $\Gamma^0 + \mathcal{R}^0\mathcal{T}_\xi$ and the impulse responses to policy shocks from $\mathcal{R}^0$ to $\mathcal{R}^0 + \mathcal{R}^0\mathcal{T}_\epsilon$, so that knowledge of the impulse responses $\Theta^0 = (\Gamma^0, \mathcal{R}^0)$ is sufficient to compute the policy rule counterfactuals embedded in the $\mathcal{T}$ adjustments.

We now define the auxiliary loss function

$$\begin{aligned} \mathsf{L}(\mathcal{T}; \phi^0, \theta) &= \mathbb{E}\mathbf{Y}'\mathcal{W}\mathbf{Y} \qquad \text{with} \quad \mathbf{Y} = (\Theta^0 + \mathcal{R}^0\mathcal{T})\mathbf{S} \\ &= \mathrm{Tr}[(\Theta^0 + \mathcal{R}^0\mathcal{T})'\mathcal{W}(\Theta^0 + \mathcal{R}^0\mathcal{T})] \,, \end{aligned} \tag{22}$$

which allows to trace out how changing the policymaker's reaction to any individual shock affects the loss. The optimal reaction adjustment is the adjustment that minimizes the auxiliary loss function, i.e. $\mathcal{T}^* = \mathrm{argmin}_{\mathcal{T}} \mathsf{L}(\mathcal{T}; \phi^0, \theta)$ with properties summarized as follows.

LEMMA 3: *Given the generic model* (14) *with* $\phi^0 \in \Phi$, *we have*

$$\mathcal{T}^* = -(\mathcal{R}^{0'}\mathcal{W}\mathcal{R}^0)^{-1}\mathcal{R}^{0'}\mathcal{W}\Theta^0 \qquad \textit{and} \qquad \mathsf{L}(\mathcal{T}^*; \phi^0, \theta) = \mathcal{L}^{\mathrm{opt}} \,. \tag{23}$$

Lemma 3 states that the auxiliary loss function, when evaluated at $\mathcal{T}^*$, attains the minimum loss. This is our identification result: knowledge of the optimal reaction to the different structural shocks is sufficient to fully characterize the optimal policy rule and to compute the minimum attainable loss $\mathcal{L}^{\mathrm{opt}}$. In addition, the lemma shows that the optimal rule adjustments can be computed using only the impulse responses $\Theta^0 = (\Gamma^0, \mathcal{R}^0)$.

Using this Lemma, we can derive our main result.

PROPOSITION 1: Given the generic model (14) with $\phi^0 \in \Phi$, we have

1. The distance to minimum loss (DML) statistic is given by

$$\Delta\mathcal{L} = \sum_{s \in \mathcal{N}} \Delta\mathcal{L}_s \,, \qquad \text{with} \qquad \Delta\mathcal{L}_s = \Theta_s^{0'}\mathcal{W}\mathcal{R}^0(\mathcal{R}^{0'}\mathcal{W}\mathcal{R}^0)^{-1}\mathcal{R}^{0'}\mathcal{W}\Theta_s^0 \,, \tag{24}$$

where $\mathcal{N}$ denotes an index set for the different shocks.

2. The optimal adjustment to the policy path response to shock $s \in \mathcal{N}$ is given by

$$\Delta\Theta_{p,s} = \mathcal{R}_p^0(\mathcal{R}^{0'}\mathcal{W}\mathcal{R}^0)^{-1}\mathcal{R}^{0'}\mathcal{W}\Theta_s^0 \,. \tag{25}$$

Proposition 1 shows how our two policy evaluation statistics —the distance to minimum loss ($\Delta\mathcal{L}$) and the optimal adjustment to the policy path response to shocks ($\Delta\Theta_{p,s}$)— can be computed using only the impulse responses under ϕ^0.

The first evaluation statistic —the distance to minimum loss— conveys how much lower the loss could have been if the policymaker had performed optimally. It is a measure of *overall* performance. Then, we can decompose the total distance $\Delta\mathcal{L}$ into shock-specific distances to minimum loss $\Delta\mathcal{L}_s$: the distance to minimum loss conditional on one type of shock only. It is the loss that could have been avoided by responding optimally to the shock s. This decomposition allows to isolate the types of shocks that were most costly to overall performance. The second evaluation statistic —the optimal adjustment to the policy path response to shocks— allows us to assess the reaction function directly: $\Delta\Theta_{p,s}$ measures how the policymaker should have adjusted her policy path response to each specific macro shock s.

4.2. *Optimal reaction adjustment under subset identification*

Proposition 1 requires the identification of all elements of $\Theta^0 = (\Gamma^0, \mathcal{R}^0)$ and $\mathcal{R}_p^0$, which in turn requires identifying all the different types of policy and non-policy shocks that can affect the economy at all horizons. In practice, this is a stringent requirement. For that reason, we now discuss how to evaluate policymakers with subset-shock identification, i.e., when the researcher can only identify a subset of the policy shocks $\boldsymbol{\epsilon} = (\epsilon_0, \epsilon_1, \ldots)'$ and non-policy shocks $\Xi = (\xi_0', \xi_1', \ldots)'$.

We denote the subsets of identified shocks by $\boldsymbol{\epsilon}_{\mathcal{S}} = \Omega_\epsilon \boldsymbol{\epsilon}$ and $\Xi_{\mathcal{S}} = \Omega_\xi \Xi$, where each Ω defines which linear combinations of shocks at different horizons are identified. Indeed, a credible identification strategy isolates exogenous variation in a variable of interest, but this variation can be a combination of news shocks at different horizons (including contemporaneous shocks). As an example, consider the narratively identified military defense spending news shocks of Ramey and Zubairy (2018); it is unlikely that this shock pertains to a unique news shock at a specific horizon h, but it is reasonable to assume that the shock is a linear combination of the different news shocks to military spending in Ξ. A similar reasoning holds for the identified policy shocks, which are likely a combination of contemporaneous and news shocks at different horizons. We let $\mathbf{S}_{\mathcal{S}} = (\Xi'_{\mathcal{S}}, \epsilon'_{\mathcal{S}})'$ be the vector of these identified shocks. The corresponding subsets of identified impulse responses are $\Theta^0_{\mathcal{S}} = \left(\Gamma^0_{\mathcal{S}}, \mathcal{R}^0_{\mathcal{S}}\right)$ and $\Theta^0_{\mathcal{S},p} = \left(\Gamma^0_{\mathcal{S},p}, \mathcal{R}^0_{\mathcal{S},p}\right)$.

In this scenario, without further assumptions, we cannot compute the entire distance to minimum loss $\Delta\mathcal{L}$, nor the entire map of policy path adjustments $\Delta\Theta_p$. However, we can construct *subset* policy evaluation statistics. To make this precise, we proceed similarly to Section 4.1 and consider the augmented policy rule

$$\mathcal{A}^0_{pp}\mathbf{P} - \mathcal{A}^0_{py}\mathbf{Y} = \mathcal{B}^0_{p\xi}\boldsymbol{\Xi} + \mathcal{B}^0_{p\epsilon}\boldsymbol{\epsilon} + \mathcal{B}^0_{p\epsilon}\Omega'_\epsilon\mathcal{T}_{\mathcal{S}}\mathbf{S}_{\mathcal{S}} \ , \tag{26}$$

where $\mathcal{T}_{\mathcal{S}}$ is a map of reaction adjustments to the identified shocks $\mathbf{S}_{\mathcal{S}}$.

Compared to (20), note how the policy rule adjustment is restricted in two ways. First, $\mathcal{T}_{\mathcal{S}}$ only adjusts the policy responses to the shocks that are identified —$\mathbf{S}$ becomes $\mathbf{S}_{\mathcal{S}}$—. Second, $\mathcal{T}_{\mathcal{S}}$ only adjusts the policy path responses as permitted by the weighting matrix Ω_ϵ —$\mathcal{B}^0_{p\epsilon}$ becomes $\mathcal{B}^0_{p\epsilon}\Omega'_\epsilon$—. In Section 4.3 we come back to these restrictions, as they are important to understand how to conduct policy evaluations using subset statistics.

As in the full identification case, we can define the policy evaluation statistics. The subset optimal reaction adjustment is defined as $\mathcal{T}^*_{\mathcal{S}} = \operatorname{argmin}_{\mathcal{T}_{\mathcal{S}}} \mathsf{L}(\mathcal{T}_{\mathcal{S}};\phi^0,\theta)$, with $\mathsf{L}(\mathcal{T}_{\mathcal{S}};\phi^0,\theta) \propto \mathrm{Tr}[(\Theta^0_{\mathcal{S}} + \mathcal{R}^0_{\mathcal{S}}\mathcal{T}_{\mathcal{S}})'\mathcal{W}(\Theta^0_{\mathcal{S}} + \mathcal{R}^0_{\mathcal{S}}\mathcal{T}_{\mathcal{S}})]$ similarly to (22). The corresponding subset distance to minimum loss is defined as $\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}} = \mathcal{L}^0 - \mathsf{L}(\mathcal{T}^*_{\mathcal{S}};\phi^0,\theta)$, which is the distance to minimum loss achievable by (i) adjusting the policy response to the subset of identified shocks and (ii) using only the policy path perturbations implied by the identified policy shocks. Similarly, the subset optimal path adjustments are given by $\Delta_{\mathcal{S}}\Theta_{\mathcal{S},p} = \Theta^0_{\mathcal{S},p} - \Theta^{\text{opt}}_{\mathcal{S},p}$ with $\Theta^0_{\mathcal{S},p} = (\Gamma^0_{\mathcal{S},p}, \mathcal{R}^0_{\mathcal{S},p})$ and $\Theta^{\text{opt}}_{\mathcal{S},p} = \Theta^0_{\mathcal{S},p} + \mathcal{R}^0_{\mathcal{S},p}\mathcal{T}^*_{\mathcal{S}}$.

PROPOSITION 2: Given the generic model (14) with $\phi^0 \in \Phi$, we have

1. The subset distance to minimum loss (DML) statistic is given by

$$\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}} = \sum_{s\in\mathcal{N}_{\mathcal{S}}} \Delta_{\mathcal{S}}\mathcal{L}_s \qquad \text{with} \qquad \Delta_{\mathcal{S}}\mathcal{L}_s = \Theta^{0'}_s\mathcal{W}\mathcal{R}^0_{\mathcal{S}}(\mathcal{R}^{0'}_{\mathcal{S}}\mathcal{W}\mathcal{R}^0_{\mathcal{S}})^{-1}\mathcal{R}^{0'}_{\mathcal{S}}\mathcal{W}\Theta^0_s \ , \tag{27}$$

where $\mathcal{N}_{\mathcal{S}}$ is an index set for the identified shocks.

2. The subset optimal adjustment to the policy path response to shock s is given by

$$\Delta_{\mathcal{S}}\Theta_{p,s} = \mathcal{R}^0_{\mathcal{S},p}(\mathcal{R}^{0\prime}_{\mathcal{S}}\mathcal{W}\mathcal{R}^0_{\mathcal{S}})^{-1}\mathcal{R}^{0\prime}_{\mathcal{S}}\mathcal{W}\Theta^0_s \, . \tag{28}$$

Proposition 2 parallels Proposition 1 in the case of subset shock identification; establishing how we can compute the "subset" policy evaluation statistics from the impulse responses to a subset of identified shocks.

4.3. *Evaluation under Subset Identification*

With Proposition 2 in hand, a researcher can evaluate policy makers based on the subset evaluation statistics, assessing performance from $\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}}$ and shock-specific performance from $\Delta_{\mathcal{S}}\mathcal{L}_s$ and $\Delta_{\mathcal{S}}\Theta_{p,s}$. However, it is important to understand how subset identification affects the computation of the minimum feasible loss and thereby how one should interpret the subset evaluation statistics. To that effect, the following lemma helps to clarify the discussion.

LEMMA 4: *The optimization problem* $\min_{\mathcal{T}} \mathsf{L}(\mathcal{T};\phi^0,\theta)$ *given* (20) *is equivalent to the unconstrained optimization problem*

$$\min_{\Delta\Theta_p} \mathrm{Tr}\left[\Theta'\mathcal{W}\Theta\right] \qquad \textit{with} \qquad \Theta = \Theta^0 + \mathcal{R}^0\mathcal{R}_p^{0^{-1}}\Delta\Theta_p \, . \tag{29}$$

The subset optimization problem $\min_{\mathcal{T}_{\mathcal{S}}} \mathsf{L}(\mathcal{T}_{\mathcal{S}};\phi^0,\theta)$ *given* (26) *is equivalent to the constrained optimization problem*

$$\min_{\Delta\Theta_{\mathcal{S},p}} \mathrm{Tr}[\Theta_{\mathcal{S}}{}'\mathcal{W}\Theta_{\mathcal{S}}] \qquad \textit{with} \qquad \Theta_{\mathcal{S}} = \Theta^0_{\mathcal{S}} + \mathcal{R}^0\mathcal{R}_p^{0^{-1}}\Delta\Theta_{\mathcal{S},p} \tag{30}$$
$$\textit{s.t.} \;\; \Delta\Theta_{\mathcal{S},p} \in \mathrm{col}(\mathcal{R}^0_{\mathcal{S},p}) \, ,$$

where $\mathrm{col}(\mathcal{R}^0_{\mathcal{S},p})$ *denotes the column space of* $\mathcal{R}^0_{\mathcal{S},p}$.

The first part of Lemma 4 states that the loss minimization with respect to $\mathcal{T}$ can be equivalently reformulated as a loss minimization problem with respect to $\Delta\Theta_p$, the vector of optimal adjustments to the policy path responses to shocks. Intuitively, characterizing the optimal rule coefficients in front of shocks is equivalent to searching for the best policy path responses to shocks. The second part states a similar result in the subset identification case. However, the subset optimization problem (30) differs from (29) in two important ways that will affect the interpretation of the evaluation statistics $\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}}$ and $\Delta_{\mathcal{S}}\Theta_{\mathcal{S},p}$.

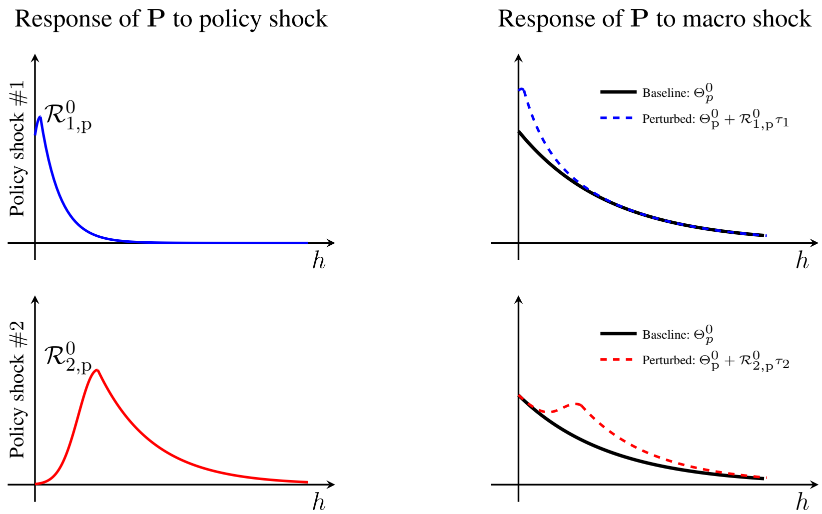
First, only the effects of the identified shocks ($\Theta_{\mathcal{S}}$) are included in the definition of the loss function in (30), so that the constructive formulation of the total DML —the DML for *all* shocks— is incomplete, being only based on the subset of shocks that could be identified. Thus, while $\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}}$ will still provide a summary measure of performance, it may not be an exhaustive measure of performance, as some shocks are missing and $\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}} \leq \Delta_{\mathcal{S}}\mathcal{L}$ where $\Delta_{\mathcal{S}}\mathcal{L} = \sum_{s\in\mathcal{N}} \Delta_{\mathcal{S}}\mathcal{L}_s$ is the *total* distance to minimum loss, the distance for *all* shocks. We will call this the "missing shocks" problem.

Second, with a limited set of policy shocks, the subset optimization is "partial" in that not all possible counterfactual policy paths are considered when searching for the optimum. We call this the "missing path" problem, and it is best understood by contrasting the optimization problems (29) and (30) in Lemma 4: when we only identify a subset of the policy shocks, the optimization problem is a *constrained* optimization problem, in that the policy path response adjustments are constrained to lie in the span of $\mathcal{R}^0_{\mathcal{S},p}$ —the responses of the policy path to the identified policy shocks—. Intuitively, the policy path response adjustment can only be a linear combination of the columns of $\mathcal{R}^0_{\mathcal{S},p}$, as these are the only policy path perturbations whose effects on the economy can be identified given the available subset of policy shocks.

A simple example helps make this policy path restriction more concrete. Figure 2 shows the impulse responses of the policy instrument to two hypothetical policy shocks that a researcher has identified. The first policy shock affects the policy path in the short term ($\mathcal{R}_{1,p}$), and the second shock affects the policy path in the medium term ($\mathcal{R}_{2,p}$). The subset optimal policy path response adjustment $\Delta\Theta_{\mathcal{S},p}$ implied by these two shocks can only be a linear combination of $\mathcal{R}_{1,p}$ and $\mathcal{R}_{2,p}$ and will thus only explore the consequences of changing the short-to-medium end of the policy path response to shocks. This "missing paths" problem does not invalidate the policy evaluation, but this is a limitation to keep in mind when interpreting the results.

In the rest of this section, we will propose approaches for "filling" these "missing shocks" and "missing paths".

FIGURE 2.—Illustration of a subset path perturbation



Note: Left column: Response of the policy path $\mathbf{P}$ to two hypothetical policy shocks: in blue (top row) or in red (bottom row). Right column: Response of the policy path $\mathbf{P}$ to a hypothetical macro shock. The response is either the baseline response under ϕ^0 ("Baseline ": Θ^0_p, black line) or after a reaction adjustment τ_1 or τ_2 ("Perturbed": $\Theta^0_p + \tau_i \mathcal{R}_{i,p}$ for $i = 1, 2$, dashed lines).

4.3.1. *Missing shocks*

Bounds on the total subset DML While the subset statistics of Proposition 2 only allow to compute the subset DML for the shocks that could be identified (the distance $\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}}$), it is possible to derive bounds for the *total* subset DML $\Delta_{\mathcal{S}}\mathcal{L}$ —the subset DML for *all* shocks— with no additional assumptions. Tight bounds imply that the missing shock problem is not severe (and the policy evaluation is close to exhaustive), whereas loose bounds imply that we might be missing a lot of relevant shocks.

To set this up consider the augmented rule $\mathcal{A}^0_{pp}\mathbf{P} - \mathcal{A}^0_{py}\mathbf{Y} = \mathcal{B}^0_{p\xi}\boldsymbol{\Xi} + \mathcal{B}^0_{p\epsilon}\boldsymbol{\epsilon} + \mathcal{B}^0_{p\epsilon}\Omega_\epsilon \widetilde{\mathcal{T}}_{\mathcal{S}}\mathbf{S}$, which is broader than the rule adjustment (26), since it adjusts the response to all shocks $\mathbf{S}$, and not just the shocks $\mathbf{S}_{\mathcal{S}}$.[6] With this we can define the total subset DML $\Delta_{\mathcal{S}}\mathcal{L} = \mathcal{L}^0 - \mathsf{L}(\widetilde{\mathcal{T}}^*_{\mathcal{S}}; \phi^0, \theta)$ which is the distance to the minimum loss defined by optimally responding to *all* shocks (in the directions allowed by the identified policy shocks). $\Delta_{\mathcal{S}}\mathcal{L}$ cannot

[6]Note that the directions of the response adjustment however remain restricted by Ω_ϵ, i.e., by the identified policy shocks. This is the missing path problem, which we turn to below.

be estimated without further assumptions, but the following corollary provides easy-to-compute bounds.

COROLLARY 1: Given the generic model (14) with $\phi^0 \in \Phi$, we can bound the total subset distance to minimum loss $\Delta_{\mathcal{S}}\mathcal{L}$ using

$$\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}} \leq \Delta_{\mathcal{S}}\mathcal{L} \leq \Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}} + \mathcal{E}^0_{\mathcal{S}} \ , \tag{31}$$

where the unexplained loss term is given by

$$\mathcal{E}^0_{\mathcal{S}} = \mathcal{L}^0 - \mathcal{L}^0_{\mathcal{S}} \qquad \text{with} \qquad \mathcal{L}^0_{\mathcal{S}} = \mathrm{Tr}(\Theta^{0'}_{\mathcal{S}} \mathcal{W} \Theta^0_{\mathcal{S}}) \ . \tag{32}$$

We note that the unexplained loss term $\mathcal{E}^0_{\mathcal{S}}$ —the loss that cannot be accounted by the subset of identified shocks— can be computed from the same set of sufficient statistics. Specifically, $\Delta_{\mathcal{S}}\mathcal{L}$ and $\mathcal{L}^0_{\mathcal{S}}$ can be measured from the sufficient statistics $\Gamma^0_{\mathcal{S}}$ and $\mathcal{R}^0_{\mathcal{S}}$, and the realized loss gives an estimate of $\mathcal{L}^0 = \mathbb{E}(\mathbf{Y}'\mathcal{W}\mathbf{Y})$ computed under ϕ^0.

Corollary 1 shows that we can use $\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}}$ to get bounds on $\Delta_{\mathcal{S}}\mathcal{L}_{\mathcal{S}}$, the subset DML for all shocks. Intuitively, it does so by exploiting the (estimable) unexplained loss $\mathcal{E}^0_{\mathcal{S}}$. The lower bound corresponds to the case where the unexplained loss is already minimal, that is could not have been lowered with another reaction function, while the upper bound corresponds to the case where the unexplained loss could have been entirely set to zero with a different reaction function.

Computing the total subset DML from reduced form innovations It is possible to estimate $\Delta_{\mathcal{S}}\mathcal{L}$, the subset distance to minimum loss for all shocks, with one additional assumption: that it is possible to compute oracle forecasts for $\mathbf{Y}$. The key is to effectively substitute structural shocks with innovations to the oracle forecast for $\mathbf{Y}$, i.e. interchanging structural shocks for reduced form shocks. To set this up, we briefly deviate from our timeless perspective and append a subscript t to $\mathbf{Y}$, i.e. $\mathbf{Y}_t = (y_t', y_{t+1}', \ldots)'$. Next, we define the innovation

$$\mathbf{U}^0_t = \mathbb{E}_t\mathbf{Y}^0_t - \mathbb{E}_{t-1}\mathbf{Y}^0_t \ , \quad \text{where} \quad \mathbb{E}_t\mathbf{Y}^0_t = \mathbb{E}(\mathbf{Y}^0_t | \mathcal{F}_t) \ , \tag{33}$$

with $\mathcal{F}_t$ the information set that includes all structural shocks are measurable up to time t and the ^{0} superscript reminds that these are innovations under ϕ^0.

COROLLARY 2: Given the generic model (14) with $\phi^0 \in \Phi$, we have that the total subset DML is given by

$$\Delta_{\mathcal{S}}\mathcal{L} = \mathrm{Tr}(\Psi^{0'}\mathcal{W}\mathcal{R}^0_{\mathcal{S}}(\mathcal{R}^{0'}_{\mathcal{S}}\mathcal{W}\mathcal{R}^0_{\mathcal{S}})^{-1}\mathcal{R}^{0'}_{\mathcal{S}}\mathcal{W}\Psi^0)\ , \tag{34}$$

where Ψ^0 is such that $\Sigma^0_U = \Psi^0\Psi^{0'}$ and $\Sigma^0_U = \mathbb{E}(\mathbf{U}^0_t\mathbf{U}^{0'}_t)$.

Corollary 2 builds on Caravello et al. (2024) who use an invertible VAR to construct the oracle innovations. The corollary shows that if a researcher is confident that she can compute oracle forecasts for $\mathbf{Y}$ (from a VAR or some other method), then it is possible to obtain an estimate for $\Delta_{\mathcal{S}}\mathcal{L}$ using the variance of the innovations to the path forecast. Intuitively, with an oracle forecast, the innovations $\mathbf{U}_t$ span all structural shocks (including the missing ones), and we can use the impulse responses to Wold innovations to compute $\Delta_{\mathcal{S}}\mathcal{L}$. While the innovations have no structural interpretation and are not uncorrelated (and thus cannot be used to decompose $\Delta_{\mathcal{S}}\mathcal{L}$ into shock-specific distances), they are sufficient to compute the total distance to minimum loss. Of course, obtaining an oracle forecast is not an innocuous task as it requires spanning the entire information set $\mathcal{F}_t$. In the context of a VAR, this requires the VAR to be invertible, see Caravello et al. (2024).

4.3.2. *Missing paths: A matrix completion problem*

We can think of the missing paths problem as a matrix completion problem where the data only provide enough variation to credibly identify a few columns or combinations of columns of the matrix needed to compute policy counterfactuals— that is, the matrix $\mathcal{R}^0$. This type of problem also appears in other strands of the macro literature, e.g. Auclert et al. (2018) with the identification of the intertemporal MPC matrix.

The $\mathcal{R}^0$ matrix completion problem is an active area of research in macro (McKay and Wolf, 2023), and we briefly discuss broad strategies that can help in our policy evaluation setting. The general strategy consists in reducing the dimension of $\mathcal{R}^0$ by imposing restrictions on the matrix $\mathcal{R}^0$, either high-level assumptions such as smoothness or explicit restrictions coming from a structural model.

Functional approximation A first strategy, in the spirit our our semi-structural approach to policy evaluation, is to reduce the matrix's dimension by imposing high-level restrictions on the shape of the impulse responses to policy shocks; for instance imposing smoothness in the effects of news shocks: while news shocks have different effects on $\mathbf{Y}$, their effects are likely to vary smoothly with the horizon of the news shock. Conceptually, this amounts to using reduced rank or functional approximations for approximating $\mathcal{R}^0$, e.g. with basis functions, such as polynomials or B-splines (e.g., Barnichon and Brownlees, 2018, Inoue and Rossi, 2021). We provide a sketch of this approach in the supplementary material Barnichon and Mesters (2025, Section S6).

Structural approximation An alternative strategy is to complete the matrix $\mathcal{R}^0$ with structural models. In particular, Caravello et al. (2024) show how a structural model can be used in a minimal way, i.e., using the structural model only to extrapolate the "missing" policy paths from the identified policy shocks. Two particularly attractive features are that the method does not require the model to explicitly specify (i) the structural shocks driving the business cycle, and (ii) the policy block of the economy —the policy rule—.

4.4. *Comparison under Subset Identification*

Consider now a setting where we have observed the terms of two policy makers A and B, and we want to compare their performance. Recall that a subset-evaluation is only able to evaluate the policy path reaction function in specific directions of improvements, i.e., based on the policy path adjustments made possible by the identified policy shocks. For the comparison to be relevant, we must first make sure that we are assessing the policy path responses of A and B in the same "directions". The following proposition provides a simple-to-verify condition under which the comparison is on equal grounds.

PROPOSITION 3: For two policy makers A and B, let $\mathcal{R}^A_{\mathcal{S},p}$ and $\mathcal{R}^B_{\mathcal{S},p}$ denote their subset policy path response maps. We have that if $\text{col}(\mathcal{R}^A_{\mathcal{S},p}) = \text{col}(\mathcal{R}^B_{\mathcal{S},p})$, then the subset-DMLs (27) for any non-policy shock s are invariant to differences in the weights Ω^A_ϵ and Ω^B_ϵ.

The proposition follows straightforwardly from Lemma 4. With only *one* identified policy shock, the condition boils down to having the policy impulse response $\mathcal{R}_{\mathcal{S},p}$ being equal across policy makers. Intuitively, when the subset identifications over two periods A and B imply the same $\mathcal{R}_{\mathcal{S},p}$, then the constraints on the optimization problem are the same for A and B, and the subset evaluation statistics will assess the same horizons of the policy path response to shocks.[7]

With that important step taken care of, two different comparisons can then be conducted: an overall comparison and a comparison by shock category.

Overall comparison

An overall comparison is based on the total subset DML $\Delta_{\mathcal{S}}\mathcal{L}$; the (subset) distance to minimum loss for all the different shocks that affected the economy during a policy maker's term. For each policymaker we can compute bounds for $\Delta_{\mathcal{S}}\mathcal{L}$ from Corollary 1, or estimate $\Delta_{\mathcal{S}}\mathcal{L}$ using Corollary 2 and the additional assumption that credible oracle forecasts can be constructed. We can then state that policy maker A performed better to B overall whenever $\Delta_{\mathcal{S}}\mathcal{L}^A < \Delta_{\mathcal{S}}\mathcal{L}^B$.

Categorical comparison

Digging deeper, a researcher may be interested in comparing performances based on the responses to any specific shock. Note that in the context of our generic model, all macro shocks are equivalent from the policy maker's perspective and no shock is more difficult to respond to than another. In practice however, policy makers may face different issues when faced with different types of shocks.[8] By conditioning a policy maker's comparison on a particular shock, we can "control" for these shock-specific practical issues.[9]

---

[7]When the impulse responses $\mathcal{R}_{\mathcal{S},p}$ do not lie in the same linear subspace, more work needs to be done, with possible solutions including the use functional or structural models to extrapolate for the unidentified shocks, as discussed above.

[8]For multiple reasons. First, certain shocks may be harder to detect in real time, for instance TFP shocks as illustrated by the 90s surge in productivity that was initially invisible in the aggregate data (e.g., Bernanke, 2022). Second, certain shocks (e.g., oil shocks) confront policy makers with more difficult trade-offs, which can expose them to political pressure (as the Fed in the 1970s, see Drechsel, 2024). Third, some shocks are less well understood than others, think of a COVID shock, a tariff shock, or even a financial shock. In the 1930s, the Fed had little understanding on the disastrous effects of bank failures on the money supply and the economy (Friedman and Schwartz, 1963, Bernanke, 1983).

[9]Digging further, one may even want to condition on a specific shock s, say a news to a financial shock that will realize in 4 quarters. This is difficult for two reasons. First, the macro data are typically not be rich enough to separately identify such a finely-defined type of shock. Second, over a policy maker's term, this particular shock

To that effect, we define Ξ_c as the sequence of shocks in category c, say financial, oil, TFP, etc. The ideal statistics for comparing A and B are then the category-specific subset-DMLs $\Delta_{\mathcal{S}}\mathcal{L}_c$, which are defined as the sum of the shock-specific subset DMLs $\Delta_{\mathcal{S}}\mathcal{L}_s$ for all shocks in category c.

Computing $\Delta_{\mathcal{S}}\mathcal{L}_c$ thus requires the identification of all shocks in category c. Otherwise, as with the missing shocks problem, the construction of $\Delta_{\mathcal{S}}\mathcal{L}_c$ is incomplete, and a comparison of performance based on category c may not be exhaustive if some shocks in category c are missing. To remedy this limitation, different solutions are possible. A narrative approach can be used to ensure sure that no important shock in category c is missing, for instance by using narrative accounts to make sure that all the key episodes of financial distress are used to construct $\Delta_{\mathcal{S}}\mathcal{L}_c$.

Alternatively, one can take a more reduced-form approach by exploiting an additional assumption like Corollary 2 —the ability to construct oracle forecasts—. Intuitively, from the forecast revisions to the category-defining variable (say oil price inflation for the oil shock category), we can compute a subset DML in which no structural shock in category c is left out and compare how well policy makers responded to forecast revisions to a particular category.[10] To set this up, let $\mathbf{U}_t^c = \mathbb{E}_t\mathbf{C}_t - \mathbb{E}_{t-1}\mathbf{C}_t$ be the forecast innovation associated with the variable $\mathbf{C}_t = (c_t, c_{t+1}, \ldots)'$ that defines the category. We can decompose the forecast innovations for $\mathbf{Y}$ using $\mathbf{U}_t = \mathbf{D}\mathbf{U}_t^c + \mathbf{U}_t^{-c}$, where $\mathbf{D}$ is the projection coefficient and $\mathbf{U}_t^{-c}$ is a remainder.

COROLLARY 3: Given the generic model (14) with $\phi^0 \in \Phi$, we have that the total subset DML can be decomposed as $\Delta_{\mathcal{S}}\mathcal{L} = \Delta_{\mathcal{S}}\mathcal{L}_c + \Delta_{\mathcal{S}}\mathcal{L}_{-c}$ with

$$\Delta_{\mathcal{S}}\mathcal{L}_{U_c} = \mathrm{Tr}(\Psi^{c'}\mathbf{Q}_{\mathcal{S}}\Psi^{c}) \qquad \text{and} \qquad \Delta_{\mathcal{S}}\mathcal{L}_{-U_c} = \mathrm{Tr}(\Psi^{-c'}\mathbf{Q}_{\mathcal{S}}\Psi^{-c})$$

may not even have realized. While we write out generic model with an infinite set of news shocks at each horizon, in practice policy makers face a much smaller set of shocks during their term.

[10]Another reduced-form approach in the spirit of Angeletos et al. (2020) would be to compare policy makers based on their reaction to a "dominant" shock in each category, that is to compare policy makers based on their reaction to the reduced-form shock that explains (for each policy maker's term) most of the business cycle variance of a category-defining variable (say energy inflation).

where $\mathbf{Q}_{\mathcal{S}} = \mathcal{W}\mathcal{R}^0_{\mathcal{S}}(\mathcal{R}^{0'}_{\mathcal{S}}\mathcal{W}\mathcal{R}^0_{\mathcal{S}})^{-1}\mathcal{R}^{0'}_{\mathcal{S}}\mathcal{W}$ and Ψ^c and Ψ^{-c} are such that $\Sigma^c_U = \Psi^c\Psi^{c'}$ and $\Sigma^{-c}_U = \Psi^{-c}\Psi^{-c'}$, with $\Sigma^c_U = \mathbf{D}\mathbb{E}(\mathbf{U}^c_t\mathbf{U}^{c'}_t)\mathbf{D}'$ and $\Sigma^{-c}_U = \mathbb{E}(\mathbf{U}^{-c}_t\mathbf{U}^{-c'}_t)$.

The corollary shows how the total subset DML can be decomposed into the contribution of category c ($\Delta_{\mathcal{S}}\mathcal{L}_{U_c}$) and a remainder term. With this we can compare A and B by computing $\Delta_{\mathcal{S}}\mathcal{L}_{U_c}$.

There are trade-offs between using $\Delta_{\mathcal{S}}\mathcal{L}_c$ and $\Delta_{\mathcal{S}}\mathcal{L}_{U_c}$. On the one hand, $\Delta_{\mathcal{S}}\mathcal{L}_c$ has a structural interpretation —it allows to compare how well each policy maker responded to structural shocks in category c, for instance oil shocks—, but it can be difficult to measure all structural shocks in a given category leaving open the possibility that missing shocks distort the comparison. On the other hand, $\Delta_{\mathcal{S}}\mathcal{L}_{U_c}$ only has a reduced form interpretation —it allows to compare how well each policy maker responded to forecast revisions to category c, for instance forecast revisions of oil inflation—, but it provides an exhaustive comparison: there is no possibility that missing shocks in category c distort the comparison.

5. EVALUATING US MONETARY POLICY, 1879-2019

In this section, we apply our methodology to evaluate the conduct of monetary policy in the US over the 1879-2019 period. We consider four distinct periods: (i) the classical Gold Standard period 1879-1912 before the creation of the Federal Reserve, (ii) the early Fed years 1913-1941, (iii) the post World War II period 1954-1984 and (iv) the post-Volcker period 1990-2019.[11]

During the classical Gold Standard period, there was no active monetary policy, but this period is instructive as a benchmark against which we can compare later Fed performances. The early Fed period starts with the founding of the Fed in 1913 and ends with the US entering the second World War. The post-war period starts in 1951 when the Fed regained some independence following the Treasury–Fed Accord (e.g. Romer and Romer, 2004a).[12]

[11] During the 1879-1912 Gold Standard period, when no formal policy institution existed, we take the three-month Treasury rate as the "policy rate" that a hypothetical central bank could have controlled. For the 1913-1941 early Fed period, we use the Fed discount rate as the policy rate. To capture the policy stance during the post WWII periods, we use the Fed funds rate as the policy rate.

[12] We exclude the period covering World War II until the Treasury-Fed accord of 1951, as the Fed was financing the war effort and lacked independence.

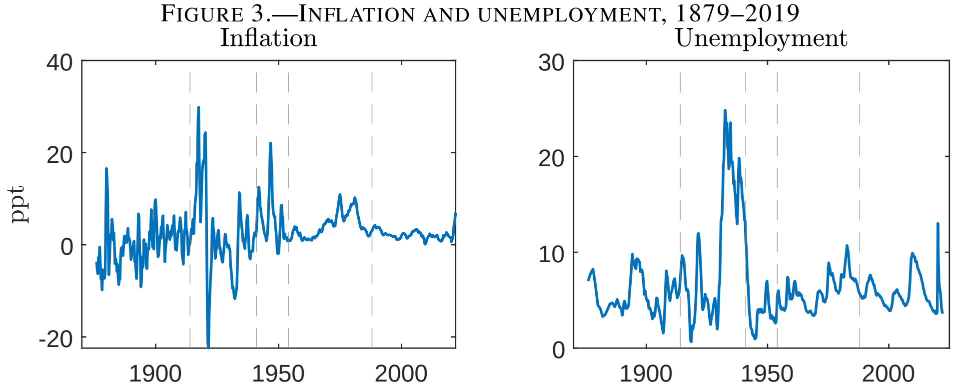
The post Volcker period covers the Great Moderation period and ends immediately before the COVID-19 pandemic.

We evaluate the Fed based on the loss function

$$\mathcal{L} = \frac{1}{2}\mathbb{E}\sum_{h=0}^{H}\beta^h(\pi_{t+h}^2 + \lambda u_{t+h}^2) \ , \tag{35}$$

where π_t denotes the inflation gap, u_t the unemployment rate gap, β the discount factor and λ the preference parameter. We set the targets $\pi^* = 2$ and $u^* = 5$, noting that the specific values of constant targets have little impact on our results. The evaluation of the reaction function is based on impulse responses to shocks —i.e., path deviations following an innovation— and therefore does not depend on the constant terms in $\mathbf{Y}$. Thus, as long as targets are constant within each period, their values are irrelevant.[13]

FIGURE 3.—INFLATION AND UNEMPLOYMENT, 1879–2019



Note: Year-on-year inflation (GDP deflator) and the unemployment rate. The vertical lines highlight the different periods: Pre Fed 1879-1912, Early Fed 1913-1941, Post WWII 1951-1984 and Post Volcker 1990-2019.

Our baseline choice for the loss function sets $\beta = \lambda = 1$, and we take $H = 40$ quarters, a horizon large enough to ensure that the impulse responses have time to mean-revert. The data are quarterly, inflation is measured as year-on-year inflation based on the output deflator from Balke and Gordon (1986), and the unemployment rate before 1948 is taken

[13]In the supplementary material, we report some robustness checks where we allow for a time-varying u^*. Results are very similar.

from the NBER Macrohistory database over 1929-1948 and extended back to 1876 by interpolating the annual series from Weir (1992) and Vernon (1994).

5.1. *Policy evaluation with sufficient statistics*

Our approach to policy evaluation requires the estimation of two sets of statistics: (i) the impulse responses of the policy objectives and the policy instrument to policy shocks, and (ii) the same impulse responses to non-policy shocks.

After detailing our shocks identifying assumptions and estimation methods, we present the results of an *overall* policy evaluation of each period. Then, to uncover the drivers of under-performance and directly evaluate the policy reaction function, we conduct *shock-specific* policy assessments.

5.2. *Shock identification*

Identification of policy shocks

We first discuss the identification of the policy shocks ϵ_S in each sub-period. Whenever possible, we draw on the state of the art in the literature and identify one monetary shock series per period. As we will see, these identified shocks generate similar policy path perturbations in each period, implying that our subset-based policy evaluation will assess policymakers in similar directions: evaluating the short end of the policy path response to shocks.[14]

For the Pre Fed Gold Standard period, there is no well established approach to identify monetary shocks, and we propose a new approach that exploits a unique feature of the Gold Standard. Under a Gold Standard, the monetary base depends on the amount of gold in circulation, which can itself vary for exogenous reasons related to the random nature of gold discoveries or development of new extraction techniques. As such, we use unanticipated large gold mine discoveries (discoveries that led to gold rushes) and the date of peak mine extraction as an instrument for movements in the monetary base. To the extent that the timing of the discovery and peak mine extraction is unrelated to the state of the business

[14]In the online appendix Section S3.1, we extend our results by identifying two monetary shock series that generate different policy path perturbations in the short-to-medium term for both the post WWII and the post Volcker periods. We obtain similar policy assessments.

cycle, gold mine discovery will be a valid instrument, see the online appendix Section S4 for more details.

For the Early Fed period, we use the Friedman and Schwartz (1963) dates extended by Romer and Romer (1989) as instruments to identify monetary policy shocks. We include five episodes —1920Q1, 1931Q3, 1933Q1, 1937Q1 and 1941Q3— where movements in money were "unusual given economic developments" (Romer and Romer, 1989). For the Post World War II period we use the Romer and Romer (2004b) monetary shocks, and for the Post Volcker period we use the high-frequency identification (HFI) approach, pioneered by Kuttner (2001) and Gürkaynak et al. (2005), and we use surprises in fed funds futures prices around FOMC announcement as proxies for monetary shocks, specifically surprises to 3-months ahead fed funds futures (FF4).[15]

Identification of non-policy shocks

We now discuss the identification of the non-policy shocks ξ_S in each sub-period. As financial shocks we use narratively identified bank panics. Each included panic was triggered by either a run on a particular trust fund or by foreign developments. The dates for the banking panics are taken from Reinhart and Rogoff (2009) and Romer and Romer (2017). To capture the severity of the bank run, each non-zero entry is rescaled by the change in the BAA-AAA spread at the time of the run, similar to the re-scaling of Bernanke et al. (1997) and Romer and Romer (2017).[16] For government spending shocks we use the news shocks to defense spending as constructed in Ramey and Zubairy (2018). As identify productivity shocks,we use the TFP series of Basu et al. (2006) over 1947-2019. To identify energy shocks, we extend the approach of Hamilton (1996) and Hamilton (2003) by identifying energy shocks as instances when energy price rises above its 3-year maximum or falls below its 3-year minimum. Since coal was the primary US energy source until World War II and oil only became the pre-dominant energy source after World War II, we measure energy price prices from the wholesale price index for fuel and lighting, available over

[15]As another robustness check we use a set-identification approach —sign restrictions as in e.g., Uhlig (2005)— for all four periods. The policy evaluation results are similar and are reported in the supplementary material.

[16]Since the time series for AAA yields only start in 1919, we backcasted AAA yields before 1919 with yields on 10-year maturity government bonds from the Macro History database (Jordà et al., 2019).

1890-2019. As measure of inflation expectations, we rely on the Livingston survey that has been continuously run over 1946-2019, and includes a question about 8-months ahead inflation expectations. Prior to World War II, there are no systematic inflation expectation survey, so we instead rely on Cecchetti (1992)'s measure of 6-months ahead inflation expectations for the Early Fed period. To identify innovations to inflation expectations, we proceed similarly to Leduc et al. (2007) and project inflation expectations on a set of controls that include past values of inflation expectation, inflation, unemployment, lags of the 3-month and 10-year treasury rates. In addition, we also project on current and past values of the other identified non-policy shocks: financial, government spending, energy price and TFP. The idea is to capture movements in inflation expectations that cannot be explained by the other shocks, i.e., that go above and beyond the typical effect of the non-policy shocks on inflation expectations.

Estimation method

To estimate impulse responses and our policy performance statistics (DML and policy path response adjustments), we rely on Bayesian structural vector autoregressive models (SVAR). For each sub-sample, we estimated a quarterly VAR with 4 lags in the variables $(\xi_{\mathcal{S},j}, \pi, u, \epsilon_{\mathcal{S}}, p)'$ where $\xi_{\mathcal{S},j}$ is the jth non-policy shock and $\epsilon_{\mathcal{S}}$ the policy shock. We order the monetary shock proxy after unemployment and inflation (and before the federal funds rate), imposing the additional restriction that monetary policy does not affect inflation and unemployment within the period, following Romer and Romer (2004b). We estimate the SVAR with Bayesian methods, which shrink the reduced form VAR coefficients using a Minnesota style prior. The prior variance hyper-parameters follow the default settings discussed in Canova and Ferroni (2025, Section 3.1.1).[17] Since the shocks are entered directly into the VAR, we compute the impulse-response functions using Cholesky identification.[18] For each draw the total subset distance to minimum loss $\Delta_{\mathcal{S}}\mathcal{L}$ (with accompanying bounds)

[17]Specifically, we set $\tau = 3$ —controlling the overall tightness—, decay $= 0.5$ —controlling the prior tightness on the lags greater than one—, $\lambda = 5$ —sum-of-coefficient prior—, $\mu = 2$ —controlling the co-persistence prior—. $\omega = 2$ —controlling the prior on the covariance matrix—.

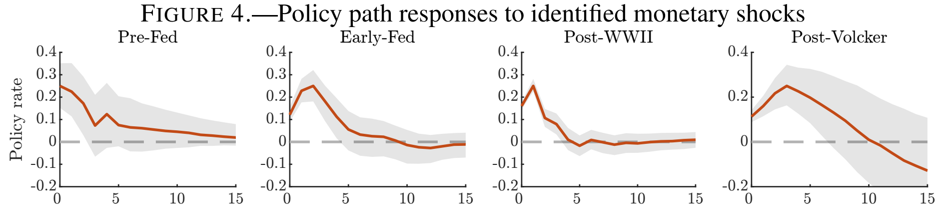
[18]Specifically, for each draw, we collect the responses of (π, u) to $\xi_{\mathcal{S},j}$ for horizons $0, 1, ..., H$ in a (1×80) vector $\Gamma^0_{\mathcal{S}}$, and the response of the policy rate p in the vector $\Gamma^0_{\mathcal{S},p}$. Similarly, the responses of (π, u, p) to $\epsilon_{\mathcal{S}}$ form a (1×80) matrix $\mathcal{R}^0_{\mathcal{S}}$ and the response of the policy rate form the vector $\mathcal{R}^0_{\mathcal{S},p}$.

and the policy path response adjustments $\Delta_{\mathcal{S}}\Theta_{p,s}$ are computed as direct functions of the impulse responses as stated in Proposition 2 and Corollaries 1-2. We report the median across draws and the 68% confidence intervals.

Policy path perturbations

Figure 4 plots the estimated responses of the policy path to each monetary shock series. The different policy shocks induce similar transitory perturbation to the short-end of the policy path (the R-squared statistics of a regression of one path on another are high and consistently above 0.65 for the first three periods), indicating that our policy evaluations will assess how well the central bank used the short-end of the policy path to respond to shocks in each period. For the Post Volcker period, the policy path perturbation is slightly more protracted than in earlier periods, so in the online appendix we study the robustness of our results by expanding the set of identified monetary shocks to capture both transitory *and* more persistent policy perturbations over the Post WWII and Post-Volcker periods. We obtained similar policy assessments.

FIGURE 4.—Policy path responses to identified monetary shocks



Note: Shaded areas depict the 68% credible sets.

5.3. *Overall evaluation*

We first report the results of our overall policy evaluation, drawing on Proposition 2 and Corollary 1 by bounding the total subset distance to minimum loss $\Delta_{\mathcal{S}}\mathcal{L}$. In addition, we estimate $\Delta_{\mathcal{S}}\mathcal{L}$ directly by using the innovations to the oracle forecasts, see Corollary 2. To do so, we estimate the innovations from a VAR that has four lags and seven variables: inflation, unemployment, the identified policy shock, energy inflation, labor productivity growth, the BAA-AAA spread, and government spending (as share of potential output). The first column of Table I reports the median estimate and 68% confidence intervals (in parentheses) for $\Delta_{\mathcal{S}}\mathcal{L}$ over the four periods. The bounds are reported inside squared brackets.

TABLE I

EVALUATION US MONETARY POLICY: 1879-2019

Panel (i) — **Distance to Minimum Loss** (DML, $\Delta_{\mathcal{S}}\mathcal{L}$)

	Overall	Shock specific					
		Bank panics	G	Energy	π^e	TFP	MP
Pre Fed 1879−1912	**8.4** (4.8,15.5) **[3,17]**	**1.5** (0.3,3.5)	**0.6** (0.1,2.1)	**0.2** (0,0.6)	—	—	**1.5** (0.7,3.2)
Early Fed 1913−1941	**97.4** (57,230) **[66,110]**	**27.7** (11.5,67.4)	**6.6** (0.8,24.3)	**1.6** (0.1,8.6)	**27.9** (9.4,70.4)	—	**18.5** (8.8,37.9)
Post WWII 1951−1984	**6.6** (3.8,12) **[3,11]**	—	**0.1** (0,0.8)	**0.9** (0.1,3.5)	**1.7** (0.3,5.5)	**0.4** (0,2.3)	**1.2** (0.4,3.2)
Post Volcker 1990−2019	**3.2** (1.4,7.2) **[1,6]**	**0.1** (0,0.6)	**0.1** (0,0.4)	**0.2** (0,0.9)	**0** (0,0.2)	**0.1** (0,0.4)	**0.7** (0.3,1.9)

Panel (ii) — **Percentage correction to policy path response** ($\frac{\text{avg}(\Delta_{\mathcal{S}}\Theta_{p,s})}{|\text{avg}(\Theta^0_{p,s})|}$)

	Shock specific					
	Bank panics	G	Energy	π^e	TFP	MP
Pre Fed 1879−1912	**−1.0** (−4.2,0)	**−0.9** (−3.1,0.1)	**−0.1** (−1.4,0.5)	—	—	—
Early Fed 1913−1941	**−2.3** (−6.2,−1)	**−1.3** (−4.4,−0.1)	**0.5** (−0.3,2.7)	**2.0** (0.8,6.7)	—	—
Post WWII 1951−1984	—	**−0.6** (−4.9,2.5)	**1.0** (0,2.1)	**4.7** (1.9,17.1)	**0.9** (−0.4,3.3)	—
Post Volcker 1990−2019	**−0.3** (−1.3,0.5)	**0.3** (−0.8,1.7)	**−0.3** (−2.8,3)	**0.1** (−1.6,1.9)	**−1.2** (−6,0.5)	—

Note: Panel (i) shows the median subset distance to minimum loss; total ($\Delta_{\mathcal{S}}\mathcal{L}$, first column) and shock-specific ($\Delta_{\mathcal{S}}\mathcal{L}_s$) together with 68% credible sets in parentheses with each row reporting estimates for a different period. In the first column, the brackets report median estimates for the upper and lower bounds for $\Delta_{\mathcal{S}}\mathcal{L}$. Panel (ii) shows the median optimal correction to the policy path reaction to each identified macro shock with 68% credible sets in parentheses. See main text for shocks definition and identification.

The Early Fed stands out as the worse performing regime with a much larger (subset) distance to minimum loss than in any other period, while the Post Volcker period stands out as the best performing one with the smallest distance to minimum loss. The Post WWII period and Pre Fed periods are relatively similar in terms of overall performance with similar distances to minimum loss. This indicates that (a) the Early Fed performed worse than the passive Gold Standard regime, and (b) the post World War II Fed still had substantial room for improvement, being only marginally superior to the Gold Standard.

This overall evaluation avoids the pitfalls of a naive approach based on the realized loss, but it does not convey the economic reasons for these different performances: it is silent about the shocks that caused sub-optimal policy performances or about possibilities for improvements. More generally, while the DML measures the loss that a superior central bank could have avoided, it does not directly assess the central bank reaction function: a distance to minimum loss could be large, because of a poor reaction function *or* because of large shocks or structural parameters that amplified the consequences of suboptimal reactions.

To answer these questions, we will turn to policy evaluations by shock category. Specifically, from Proposition 2 we will compute both the shock-specific DML $\Delta_{\mathcal{S}}\mathcal{L}_s$ *and* $\Delta_{\mathcal{S}}\Theta_{p,s}$, the optimal adjustment to the policy path response to shock s. As we saw, these statistics are the two sides of the same performance coin, and they will allow to separate the roles of good policy (appropriate reaction function) and good luck (a stable environment that limited the consequences of a sub-optimal reaction function).

5.4. *Shock-specific evaluation*

Our shock-specific evaluation considers five categories of non-policy shocks for each of the four periods (whenever possible): financial shocks, government spending shocks, energy price shocks, inflation expectations shocks, and TFP shocks. In addition, we will report the contribution of monetary shocks to policy performance, i.e., when the monetary authority itself destabilizes the economy with policy shocks.

Table I reports our results, grouping our shock-specific evaluation statistics (distance to minimum loss and optimal percentage correction) by shock category. As discussed in Section 4.4, we caution that category-specific comparisons of policy performance across periods (i.e., across rows) are only exhaustive when the identified shocks in a particular category capture all (or most) of the shocks in that category. In the supplementary material, we thus also show results based on using corollary 3 to construct $\Delta_{\mathcal{S}}\mathcal{L}_{U_c}$ and compare the Fed across periods from how well the institution responded to *forecast revisions* in each category. The policy comparisons are similar.

Results

Panel (i) of Table I reports the median and 68% confidence intervals for the (subset) distance to minimum loss for each shock s: $\Delta_{\mathcal{S}}\mathcal{L}_s$. To understand the reasons for sub-optimal performances and assess the reaction function directly, panel (ii) focuses on the optimal adjustment to the policy path response to each shock s, and specifically reports the first-year average correction to the policy path response for each broad shock type: $\frac{\text{avg}(\Delta_{\mathcal{S}}\Theta_{p,s})}{|\text{avg}(\Theta^0_{p,s})|}$. This "percentage correction" is a unitless summary measure that can convey how far off was the Fed to best responding to a particular shock type that it faced. In addition, we plot the original $\Theta^0_{p,s}$ and adjusted policy paths responses $\Theta^{\text{opt}}_{p,s}$ in the figures below. This provides a direct way to visualize the magnitude and dynamic profile of the policy path correction, i.e., the magnitude and dynamic profile of the policy mistake.

We summarize our main results below and provide more discussion and robustness checks in the online-appendix.

Improved policy in the Post Volcker period Our shock-specific evaluations confirm our overall evaluation results: we estimate strong improvements in the conduct of monetary policy in the last 30 years, i.e., roughly after Volcker's dis-inflation program. In particular, the policy path corrections are substantially smaller (and non-significant) than in the other periods, with an average (absolute) percentage correction of 30 percent post Volcker but above 100 percent in all other periods.[19]

[19] The only exception is the Post Volcker Fed's response to TFP shocks which is non-significant but larger at 120 percent. The reason is mainly because the baseline response is close to zero, in line with Gali (1999) that policy did not react to TFP shocks post 1985. That said, the welfare consequences are minimal (the TFP DML is tiny at 0.1), also in line with Gali (1999)'s conclusion that TFP shocks account for little of business cycle fluctuations.

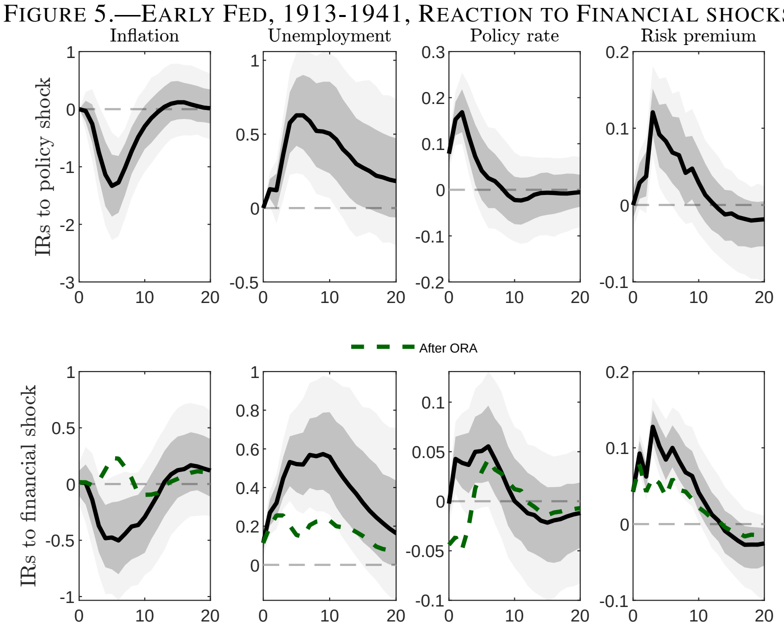
The conduct of policy did improve before Volcker however in one dimension: the variance of policy shocks went down substantially after WW II. In other words, since World War II the Fed has not directly contributed to destabilizing the economy with policy shocks. This can be seen from the DML for policy shocks, which is much smaller for the post World War II than for the early Fed (see column "MP"). In other words, even though the reaction function is not substantially superior in the post WWII period, erratic behavior in the conduct of policy was much improved after WWII, in contrast to the stop-and-go policies of the 30s or the over-reaction of the early 20s (e.g., Friedman and Schwartz, 1963, Romer, 1992).

Responding to financial shocks Focusing on the reaction to financial shocks, we can contrast the Post Volcker Fed with the Early Fed of the 1920s-1930s. Our results confirm Bernanke's promise to not repeat the mistakes of the past: the "poor" reaction function of the early Fed led to massive welfare losses, while the "good" reaction function of the Post Volcker Fed ensured little welfare losses coming from a sub-optimal reaction function.

To see this, we can first contrast the optimal policy path response adjustment for financial shocks estimated for the Early Fed period with the policy path response adjustment for the Post Volcker period. With $100 \times \frac{\mathrm{avg}(\Delta_{\mathcal{S}}\Theta_{p,s})}{|\mathrm{avg}(\Theta^0_{p,s})|} = -230\%$ (statistically significant) for the Early Fed, the Fed reaction to banking panics was much too tight; not only was the response too timid, it had the wrong sign (the adjustment being over -100%) and the policy path adjustment flips the sign of the policy path response; a decline in the policy rate instead of an increase.

To better appreciate this reaction function improvement, Figures 5 and 6 display the impulse responses underlying our calculations, for the 1913-1941 estimates and for the 1990-2019 estimates. The top rows show the impulse responses of inflation, unemployment and the interest rate to a monetary policy shock —an adverse policy shock lowering inflation and raising unemployment—, while the bottom rows show the responses of the same variables to a financial shock —an adverse financial shock lowering inflation and raising unemployment. Notice how, in the Early Fed period, the Fed *raised* the discount rate in response to adverse financial shocks. Combined with the decline in inflation caused by the financial shock, this means that the real policy rate increased substantially and monetary policy was contractionary in the aftermaths of banking panics. The (substantial) policy path response adjustment corrects this sub-optimal reaction function and turns the table on monetary policy by running an expansionary policy. After reaction adjustment (dashed green line), the policy rate goes down substantially on impact, and the paths of inflation and unemployment are consequently much more stable.

FIGURE 5.—EARLY FED, 1913-1941, REACTION TO FINANCIAL SHOCKS



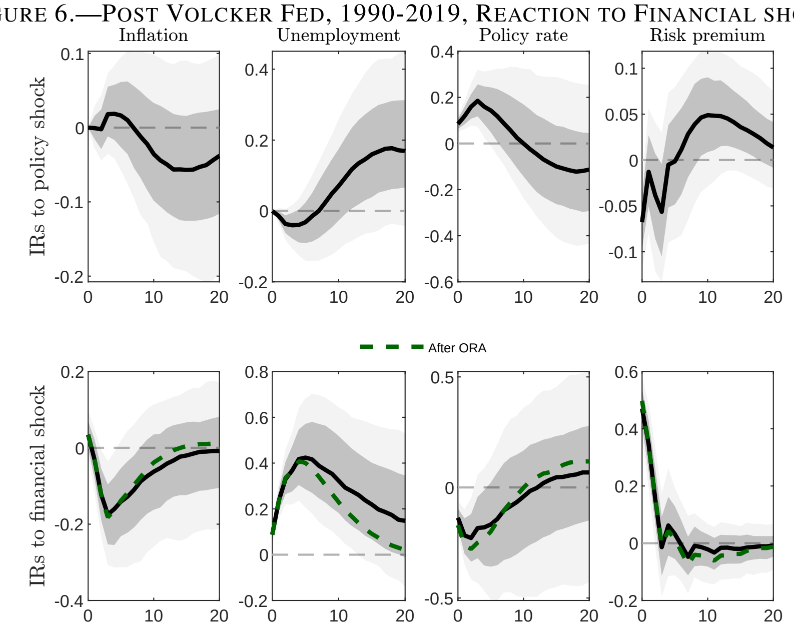
Note: The top (resp. bottom) row shows the median responses (thick line) of inflation, unemployment and the Fed's discount rate to a monetary policy shock (resp. financial shock). The dotted green lines show the adjusted impulse responses: $\Gamma^0 + \mathcal{R}^0\tau_0^*$. The 95% and 67% credible sets are plotted as dark and light shaded areas, respectively.

Consider now the evaluation of the post Volcker Fed. The estimated policy adjustment is much smaller and non significant, representing only a -30 percent correction to the original policy path response, indicating that the post Volcker Fed period reacted much more appropriately to financial shocks.[20] As a result, while the 2007-2008 financial disruptions

[20]That said, a negative path adjustment indicates that the Fed should have lowered the fed funds rate more in response to financial shocks (according to the posterior mean). This could indicate that the presence of the zero lower bound may have limited somewhat the Fed's ability to best react to the 2007-2008 financial crisis. In the supplementary material Barnichon and Mesters (2025), we generalize our method to measuring performance in a constrained optimization environment and impose a zero lower bound on interest rates. The optimal reaction adjustment is then close to zero.

were substantial, the corresponding estimated DML over the post-Volcker period is tiny (0.1); two orders magnitude smaller than the DML for financial shocks for the Early Fed (27.7). This result can also be seen in the impulse responses estimated for the Post Volcker period (Figure 6). Following a financial shock, the policy rate is negative —monetary policy is expansionary (black line, lower-right panel)—, and the policy path response adjustment (green line) is minor, leading to modest adjustments to the responses of inflation and unemployment. This is the sign of good policy.

FIGURE 6.—POST VOLCKER FED, 1990-2019, REACTION TO FINANCIAL SHOCKS



Note: The top (resp. bottom) row shows the median responses (thick line) of inflation, unemployment and the fed funds rate to a monetary policy shock (resp. financial shock). The dotted green lines show the adjusted impulse responses: $\Gamma^0 + \mathcal{R}^0\tau^*$. The 95% and 68% credible sets are plotted as dark and light shaded areas, respectively.

The early Fed vs the passive Gold Standard An interesting, and perhaps surprising, finding is that the Early Fed performed *worse* than the passive Gold standard: Both the DMLs and the policy path adjustments are larger for the Early Fed than for the Gold Standard. The total DML is about 10 times larger in the Early Fed period than in the Pre Fed period. While this difference reflects the fact that the Early Fed faced the larger shocks of the Great Depression, panel (ii) shows that the reaction function of the Early Fed is actually substantially worse across all shocks. The most striking difference is the reaction to banking panics with a correction of -230 percent for the Early Fed but "only" -100 percent for the passive Gold Standard. Interestingly, this is consistent with narrative evidence that the Early Fed was deeply influenced by the real bills doctrine and excessively worried about speculations (e.g. Humphrey and Timberlake, 2019).

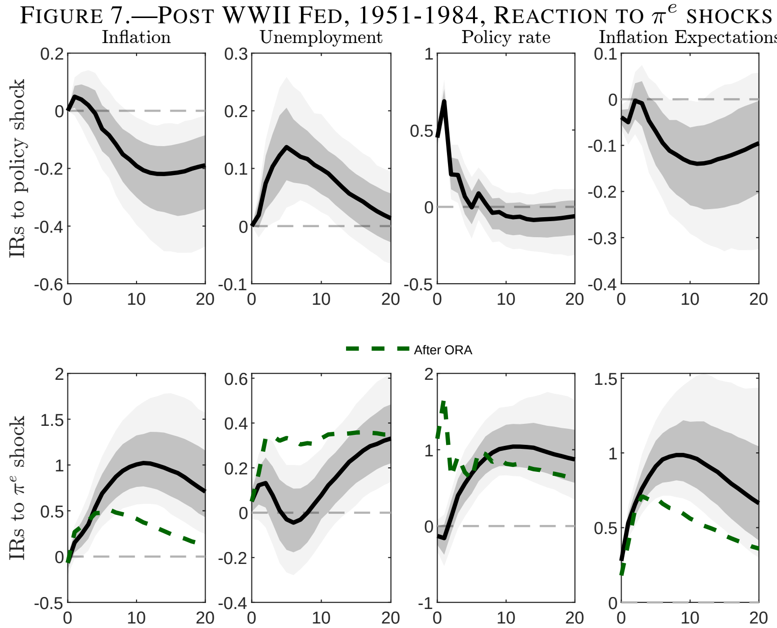
The Great Inflation US monetary policy during the 1970s has generally been considered poor (e.g., Romer and Romer, 2004a), in particular not responding more than one-to-one with changes in inflation (Clarida et al., 2000) and violating the so-called Taylor principle. However, beyond that Taylor principle, it has been difficult to quantify how "poor" monetary policy had been.

Overall, we find that the Fed's reaction function during the 60s-70s is on a par (i.e., "just as bad") with the reaction function of the early Fed, with policy path corrections of similar magnitudes, though the nature of the underlying shocks is different. Post World War II, the Fed reaction was too weak following all the different supply-type shocks that we identified: energy price shocks, TFP shocks as well as inflation expectation shocks. In fact, the reaction to inflation expectation shocks over the 60s-70s displays the largest deviation from optimality over the entire 150 year of monetary history. The consequences of these sub-optimal reactions were much smaller however, with DMLs an order of magnitude smaller for the Post WWII period than for the Early Fed period.[21]

To illustrate these sub-optimal reaction functions, Figure 7 plots the estimated impulse responses for inflation expectation shocks. In response to an inflation expectation shock, inflation rises progressively, but the policy rate does not respond, leading to negative real interest rates and further increasing inflation. The (large) policy path correction restores the Taylor principle: after correction, the policy rate rises strongly following an inflation expectation shock (lower-right panel, Figure 7) and limits the rise in inflation (at the cost of higher unemployment).

[21]There are two possible reasons why a given suboptimal reaction function translates into larger losses, i.e., into larger distances to minimum loss: (i) the magnitudes of the shocks themselves were larger in the Early Fed period, and/or (ii) the Early Fed economy was less resilient in the face of adverse disturbances than the post WWII economy.

FIGURE 7.—POST WWII FED, 1951-1984, REACTION TO π^e SHOCKS



Note: The top (resp. bottom) row shows the median responses (thick line) of inflation, unemployment and the fed funds rate to a monetary policy shock (resp. inflation expectations shock). The dotted green lines show the adjusted impulse responses: $\Gamma^0 + \mathcal{R}^0\tau^*$. The 95% and 68% credible sets are plotted as dark and light shaded areas, respectively.

6. CONCLUSION

In this paper, we propose a semi-structural method to evaluate policy makers with minimal assumptions on the underlying economic model. Specifically, for a large class of linear forward looking macro models and quadratic loss functions, it is possible to measure the distance to the optimal reaction function and distance to minimum loss from well known and estimable sufficient statistics: the impulse responses to policy and non-policy shocks.

An important open question for US monetary policy going forward is why did large and uniform improvements happen only in the last 30 years. A better understanding of the functioning of the economy (Friedman and Schwartz, 1963), better and more timely data (Romer, 1986, Orphanides, 2001), better forecasting (Dominguez et al., 1988) and better causal inference methods (Romer and Romer, 1989) could all be part of the improvements in policy over the last 30 years. Parsing out these different reasons is an important question for future research.

Our proposed methodology could be applied to many other important evaluation questions; in the context of monetary policy (e.g., comparing central banks such as the Fed vs the ECB during the Great Recession), in the context of fiscal policy (e.g., comparing the performance of US presidents, health policy (e.g., comparing governments' policy responses to COVID), or climate change mitigation policy. We leave these questions for future research.

Last, we note that our approach focuses on evaluating reaction functions given a set of objectives, but it is silent about the policy objectives themselves. For instance, what set of policy objectives should a policy institution target? And what long-run level to target? For instance, what are the costs and benefits of a change in the inflation target? Developing sufficient statistics methods to study these long-run questions would be a fruitful avenue for future research.

BIBLIOGRAPHY

ALMUNIA, MIGUEL, AGUSTIN BENETRIX, BARRY EICHENGREEN, KEVIN H O'ROURKE, AND GISELA RUA (2010): "From great depression to great credit crisis: similarities, differences and lessons," *Economic policy*, 25 (62), 219–265. [4]

ANGELETOS, GEORGE-MARIOS, FABRICE COLLARD, AND HARRIS DELLAS (2020): "Business-cycle anatomy," *American Economic Review*, 110 (10), 3030–3070. [27]

AUCLERT, ADRIEN, BENCE BARDÓCZY, MATTHEW ROGNLIE, AND LUDWIG STRAUB (2021): "Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models," *Econometrica*, 89 (5), 2375–2408. [13]

AUCLERT, ADRIEN, MATTHEW ROGNLIE, AND LUDWIG STRAUB (2018): "The intertemporal keynesian cross," Tech. rep., National Bureau of Economic Research. [24]

BALKE, N. AND R. J. GORDON (1986): *Appendix B Historical Data*, National Bureau of Economic Research. [29]

BARNICHON, REGIS AND CHRISTIAN BROWNLEES (2018): "Impulse Response Estimation By Smooth Local Projections," *Review of Economics and Statistics*. [25]

BARNICHON, REGIS AND GEERT MESTERS (2023): "A Sufficient Statistics Approach for Macroeconomic Policy," *American Economic Review*, 113 (11), 2809–45. [6]

——— (2025): "Additional supplementary material for: Evaluating Policy Institutions — 150 Years of US Monetary Policy —," Supplementary Material. [25, 38]

BASU, SUSANTO, JOHN G. FERNALD, AND MILES S. KIMBALL (2006): "Are Technology Improvements Contractionary?" *American Economic Review*, 96 (5), 1418–1448. [31]

BERNANKE, BEN S. (1983): "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *The American Economic Review*, 73 (3), 257–276. [26]

——— (1986): "Alternative explanations of the money-income correlation," *Carnegie-Rochester Conference Series on Public Policy*, 25, 49 – 99. [14]

Bernanke, Ben S (2002): "On Milton Friedman's ninetieth birthday," in *remarks at the conference to honor Milton Friedman, University of Chicago*. [4]

——— (2022): *21st century monetary policy: The Federal Reserve from the great inflation to COVID-19*, WW Norton & Company. [26]

Bernanke, Ben S, Mark Gertler, and Mark Watson (1997): "Systematic monetary policy and the effects of oil price shocks," *Brookings papers on economic activity*, 1997 (1), 91–157. [6, 31]

Blanchard, Olivier and Jordi Galí (2007): "The Macroeconomic Effects of Oil Price Shocks: Why Are the 2000s so Different from the 1970s?" in *International Dimensions of Monetary Policy*, National Bureau of Economic Research, Inc, 373–421. [5]

Blinder, Alan S. and Mark W. Watson (2016): "Presidents and the US Economy: An Econometric Exploration," *American Economic Review*, 106 (4), 1015–45. [5]

Bordo, Michael D and Finn E Kydland (1995): "The gold standard as a rule: An essay in exploration," *Explorations in Economic History*, 32 (4), 423–464. [4]

Canova, Fabio and Filippo Ferroni (2025): "A hitchhiker guide to empirical macro models," Technical Report. [32]

Caravello, Tom, Alisdair McKay, and Christian Wolf (2024): "Evaluating Policy Counterfactuals: A "VAR-Plus" Approach," . [6, 24, 25]

Cecchetti, Stephen G (1992): "Prices During the Great Depression: Was the Deflation of 1930-1932 Really Unanticipated?" *The American Economic Review*, 141–156. [32]

Clarida, Richard, Jordi Galí, and Mark Gertler (2000): "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *The Quarterly Journal of Economics*, 115 (1), 147–180. [4, 5, 40]

Coibion, Olivier and Yuriy Gorodnichenko (2011): "Monetary policy, trend inflation, and the great moderation: An alternative interpretation," *American Economic Review*, 101 (1), 341–370. [5]

Dominguez, Kathryn M, Ray C Fair, and Matthew D Shapiro (1988): "Forecasting the depression: Harvard versus Yale," *The American Economic Review*, 595–612. [41]

Drechsel, Thomas (2024): "Estimating the effects of political pressure on the fed: a narrative approach with new data," Tech. rep., National Bureau of Economic Research. [26]

Fair, Ray C. (1978): "The Use of Optimal Control Techniques to Measure Economic Performance," *International Economic Review*, 19 (2), 289–309. [5]

Friedman, Milton and Anna Jacobson Schwartz (1963): *A monetary history of the United States, 1867-1960*, Princeton University Press. [6, 26, 31, 37, 41]

Gali, Jordi (1999): "Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations?" *American economic review*, 89 (1), 249–271. [36]

Galí, Jordi (2015): *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework and its Applications*, Princeton University Press. [8, 10]

Galí, Jordi and Luca Gambetti (2009): "On the sources of the great moderation," *American Economic Journal: Macroeconomics*, 1 (1), 26–57. [4]

GALI, JORDI AND MARK GERTLER (2007): "Macroeconomic Modeling for Monetary Policy Evaluation," *Journal of Economic Perspectives*, 21 (4), 25–46. [5]

GALI, JORDI, J DAVID LÓPEZ-SALIDO, AND JAVIER VALLÉS (2003): "Technology shocks and monetary policy: assessing the Fed's performance," *Journal of Monetary Economics*, 50 (4), 723–743. [4, 5]

GÜRKAYNAK, REFET, S., BRIAN SACK, AND ERIC SWANSON (2005): "The Sensitivity of Long-Term Interest Rates to Economic News: Evidence and Implications for Macroeconomic Models," *American Economic Review*, 95, 425–436. [31]

HAMILTON, JAMES D. (1996): "This is what happened to the oil price-macroeconomy relationship," *Journal of Monetary Economics*, 38 (2), 215–220. [31]

——— (2003): "What is an oil shock?" *Journal of Econometrics*, 113 (2), 363–398. [31]

HUMPHREY, THOMAS M AND RICHARD H TIMBERLAKE (2019): *Gold, the real bills doctrine, and the Fed: Sources of monetary disorder, 1922-1938*, Cato Institute Washington, DC. [40]

INOUE, ATSUSHI AND BARBARA ROSSI (2021): "A new approach to measuring economic policy shocks, with an application to conventional and unconventional monetary policy," *Quantitative Economics*, 12 (4), 1085–1138. [25]

JORDÀ, ÒSCAR, KATHARINA KNOLL, DMITRY KUVSHINOV, MORITZ SCHULARICK, AND ALAN M TAYLOR (2019): "The rate of return on everything, 1870–2015," *The Quarterly Journal of Economics*, 134 (3), 1225–1298. [31]

KUTTNER, KENNETH N. (2001): "Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market," *Journal of Monetary Economics*, 47 (3), 523–544. [31]

LEDUC, SYLVAIN, KEITH SILL, AND TOM STARK (2007): "Self-fulfilling expectations and the inflation of the 1970s: Evidence from the Livingston Survey," *Journal of Monetary economics*, 54 (2), 433–459. [32]

LEEPER, ERIC M. AND TAO ZHA (2003): "Modest policy interventions," *Journal of Monetary Economics*, 50 (8), 1673–1700. [6]

MCKAY, ALISDAIR AND CHRISTIAN WOLF (2023): "What Can Time-Series Regressions Tell Us About Policy Counterfactuals?" *Econometrica*, forthcoming. [6, 24]

MELTZER, ALLAN H (2009): "A History of the Federal Reserve, Volumes 1-2, 1913-1986," . [6]

MISHKIN, FREDERIC S. (2010): "Will Monetary Policy Become More of a Science?" in *The Science and Practice of Monetary Policy Today*, ed. by Wieland V., Springer, Berlin, 81–103. [5]

ORPHANIDES, ATHANASIOS (2001): "Monetary policy rules based on real-time data," *American Economic Review*, 91 (4), 964–985. [41]

——— (2003): "Historical monetary policy analysis and the Taylor rule," *Journal of monetary economics*, 50 (5), 983–1022. [5]

RAMEY, VALERIE (2016): "Macroeconomic Shocks and Their Propagation," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and H. Uhlig, Amsterdam, North Holland: Elsevier. [14]

RAMEY, VALERIE A. AND SARAH ZUBAIRY (2018): "Government Spending Multipliers in Good Times and in Bad: Evidence from U.S. Historical Data," *Journal of Political Economy*, 126. [18, 31]

REINHART, CARMEN M AND KENNETH S ROGOFF (2009): "This time is different," in *This Time Is Different*, princeton university press. [31]

ROMER, CHRISTINA D (1986): "Is the Stabilization of the Postwar Economy a Figment of the Data?" *The American Economic Review*, 76 (3), 314–334. [41]

——— (1992): "What ended the great depression?" *The Journal of Economic History*, 52 (4), 757–784. [37]

ROMER, CHRISTINA D AND DAVID H ROMER (1989): "Does monetary policy matter? A new test in the spirit of Friedman and Schwartz," *NBER macroeconomics annual*, 4, 121–170. [31, 41]

——— (2004a): "Choosing the Federal Reserve chair: lessons from history," *Journal of Economic Perspectives*, 18 (1), 129–162. [28, 40]

ROMER, CHRISTINA D. AND DAVID H. ROMER (2004b): "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 94, 1055–1084. [31, 32]

ROMER, CHRISTINA D AND DAVID H ROMER (2013): "The missing transmission mechanism in the monetary explanation of the Great Depression," *American Economic Review*, 103 (3), 66–72. [4]

ROMER, CHRISTINA D. AND DAVID H. ROMER (2017): "New Evidence on the Aftermath of Financial Crises in Advanced Countries," *American Economic Review*, 107 (10), 3072–3118. [31]

SIMS, CHRISTOPHER A AND TAO ZHA (2006): "Does monetary policy generate recessions?" *Macroeconomic Dynamics*, 10 (2), 231–272. [6]

SVENSSON, LARS EO (2003): "What is wrong with Taylor rules? Using judgment in monetary policy through targeting rules," *Journal of Economic Literature*, 41 (2), 426–477. [5]

TAYLOR, JOHN B (1999): "A historical analysis of monetary policy rules," in *Monetary policy rules*, University of Chicago Press, 319–348. [5]

UHLIG, HARALD (2005): "What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure," *Journal of Monetary Economics*, 52 (2), 381–419. [31]

VERNON, J.R. (1994): "Unemployment rates in postbellum America: 1869–1899," *Journal of Macroeconomics*, 16 (4), 701–714. [30]

WEIR, DAVID R (1992): "A century of US unemployment, 1890-1990: revised estimates and evidence for stabilization," *Research in Economic History*, 14 (1), 301–46. [30]

WHEELOCK, DAVID C ET AL. (2010): "Lessons learned? Comparing the Federal Reserve's responses to the crises of 1929-1933 and 2007-2009," *Federal Reserve Bank of St. Louis Review*, 92 (Mar), 89–108. [4]