Instrument-Based vs. Target-Based Rules*

Marina Halac† Pierre Yared‡

April 15, 2020

Abstract

We study rules based on instruments vs. targets. Our application is a New Keynesian economy where the central bank has non-contractible information about aggregate demand shocks and cannot commit to policy. Incentives are provided to the central bank via punishment which is socially costly. Instrument-based rules condition incentives on the central bank's observable choice of policy, whereas target-based rules condition incentives on the outcomes of policy, such as inflation, which depend on both the policy choice and realized shocks. We show that the optimal rule within each class takes a threshold form, imposing the worst punishment upon violation. Target-based rules dominate instrument-based rules if and only if the central bank's information is sufficiently precise, and they are relatively more attractive the less severe the central bank's commitment problem. The optimal unconstrained rule relaxes the instrument threshold whenever the target threshold is satisfied.

Keywords: Policy Rules, Private Information, Delegation, Mechanism Design, Monetary Policy, Policy Objectives

JEL Classification: D02, D82, E52, E58, E61

*We would like to thank Hassan Afrouzi, V. V. Chari, John Cochrane, John Geanakoplos, Johannes Hörner, Navin Kartik, Narayana Kocherlakota, Jen Jen La'O, Alessandro Lizzeri, George Mailath, Rick Mishkin, Ken Rogoff, Larry Samuelson, and Mike Woodford for comments. Jan Knoepfle and Angus Lewis provided excellent research assistance.

†Yale University and CEPR: marina.halac@yale.edu.

‡Columbia University and NBER: pyared@columbia.edu.

1 Introduction

The question of whether to base incentives on agents' actions or the outcomes of these actions arises in various contexts. In particular, scholars and policymakers have long debated on the merits of using instrument-based vs. target-based rules to guide monetary policy.[1] Instrument-based rules evaluate central bank performance on the choice of policy, taking a number of forms such as money growth rules, interest rate rules, and exchange rate rules. In contrast, target-based rules evaluate central bank performance on the outcomes of policy, such as the realized inflation level, price level, and output growth.

Target-based rules have become increasingly popular over the last decades, with 60 central banks around the world adopting inflation targeting since 1990 (International Monetary Fund, 2019).[2] Many of the early adopters were central banks in advanced economies, like those in New Zealand, Canada, and the United Kingdom, while more recent adopters included emerging economies such as India and Russia. Nevertheless, many developing economies still guide monetary policy based only on instruments; for example, many countries in Sub-Saharan Africa remain under an instrument-based rule in the form of a fixed exchange rate regime.[3] As for the US, a number of studies describe US monetary policy as adhering to an instrument-based rule in the form of a Taylor rule from the 1980s through the early 2000s, while pursuing an inflation target since that time.[4]

In this paper, we develop a simple model to study and compare instrument-based and target-based rules. Our focus is on the implications of these rules for optimal incentives. While both rule classes are broadly used and discussed in the context of monetary policy, as just described, the theoretical literature on central bank incentives—and

[1]This debate includes many arguments for and against Taylor rules; see Svensson (2003), Bernanke (2004, 2015), McCallum and Nelson (2005), Appelbaum (2014), Blinder (2014), Taylor (2014), Kocherlakota (2016), and the Statement on Policy Rules Legislation signed by a number of economists, available at `http://web.stanford.edu/~johntayl/2016_pdfs/Statement_on_Policy_Rules_Legislation_2-29-2016.pdf`. There are also numerous discussions of inflation targeting; see Bernanke and Mishkin (1997), Bernanke et al. (1999), Mishkin (1999, 2017), Svensson (2003), and Woodford (2007), among others.

[2]This includes the 19 countries in the Eurozone plus 41 other countries classified by International Monetary Fund (2019) as inflation targeters.

[3]Of the 48 countries in sub-Saharan Africa, 21 have conventional pegs, 18 have managed exchange rates, and the rest have floating exchange rates. See International Monetary Fund (2019).

[4]See Kahn (2012), Taylor (2012), and Nikolsko-Rzhevskyy, Papell, and Prodan (2019) for the first time period and Bernanke (2012), Yellen (2012), and Shapiro and Wilson (2019) for the second one. The International Monetary Fund does not provide a classification of US monetary policy.

incentives in principal-agent relationships, more generally—has largely focused on one rule class or the other, lacking a general comparison that can inform this discussion. Using a mechanism design approach, we present a theory that elucidates the relative benefits of instrument-based vs. target-based rules and shows which class will be preferred as a function of the environment. Additionally, we characterize the optimal unconstrained or *hybrid* rule and examine how combining instruments and targets can improve welfare.

Our model builds on a canonical New Keynesian framework. Society delegates monetary policy to a central bank that lacks commitment and is therefore biased towards looser monetary policy and higher inflation at the time of choosing policy. The central bank's policy choice is observable, but optimal policy depends on non-contractible information that the central bank possesses. Specifically, the central bank observes a private forecast of aggregate demand shocks to the economy (e.g. Romer and Romer, 2000), with a more negative forecast implying a looser socially optimal choice of monetary policy. Inflation and the output gap depend on both the central bank's policy choice and the realized aggregate demand shock.

Incentives are provided to the central bank via the threat of punishment. Punishment imposes a cost on both the central bank and society, stemming for example from formal sanctions, leadership replacement, or loss of central bank budget, autonomy, or reputation.[5] We distinguish between different classes of rules depending on how incentives are structured: we say that the rule is instrument-based if punishments depend only on the central bank's policy, and the rule is target-based if punishments depend only on the outcome of policy, namely realized inflation.[6] Under either class of rule, punishments could be tailored so as to incentivize the central bank to choose the socially optimal policy, thus eliminating any distortions arising from the central bank's inflation bias. However, because punishing the central bank is socially costly, society must trade off the benefit of mitigating its inflation bias with the benefit of

[5]See Rogoff (1985), Garfinkel and Oh (1993), and Walsh (1995, 2002) for a discussion of the different forms these punishments can take. In New Zealand the Governor of the Reserve Bank is subject to dismissal if he fails to achieve the inflation target (Hüfner, 2004), and in several countries central-bank officials are subject to scrutiny of their policies with requirements of written explanations, or public hearings in the Parliament, or reviews by independent experts (Svensson, 2010). Consequences such as dismissal imply a cost not only to the central banker but to society at large.

[6]Our setting is one where the "divine coincidence" holds (e.g. Blanchard and Galí, 2007), implying a one-to-one mapping between inflation and the output gap (holding fixed expectations of the future). Therefore, a target rule based on inflation is equivalent to one based on the output gap, and giving the central bank full discretion would be optimal absent an inflation bias.

reducing equilibrium punishments.[7]

Our analysis begins by showing that, within each class, an optimal rule takes a threshold form, with violation of the threshold leading to the worst feasible punishment. In the case of an optimal instrument-based rule, the central bank is allowed to choose any policy up to a threshold and is maximally punished for choosing looser policy. The idea is analogous to that in other models of delegation (reviewed below); since the central bank is inflation-biased relative to society, this punishment structure is optimal to deter the central bank from pursuing excessively expansionary policies. In the case of an optimal target-based rule, the central bank is subject to an inflation threshold, being maximally punished if realized inflation is above it. This punishment structure also incentivizes the central bank to not choose excessively expansionary policy, since this results in higher inflation in expectation. High-powered incentives of this form arise in moral hazard settings with hidden action.[8]

Our main result uses this characterization of the optimal rules for each class to compare their performance. We show that target-based rules dominate instrument-based rules if and only if the central bank's private information is sufficiently precise. To illustrate, suppose the central bank has a perfect forecast of aggregate demand shocks. Then society guarantees the socially optimal policy by providing steep incentives under a target-based rule, where punishments do not occur *on path* because the perfectly informed central bank chooses the policy that delivers the inflation target outcome. This target-based rule strictly dominates any instrument-based rule, as the latter cannot incentivize the central bank while giving it enough flexibility to respond to its information. At the other extreme, suppose the central bank has no private information. Then society guarantees the ex-ante socially optimal policy with an instrument-based rule that ties the hands of the central bank, namely that punishes the central bank if any looser policy is chosen. This instrument-based rule strictly dominates any target-based rule, as the latter gives the central bank unnecessary discretion and requires on-path punishments to provide incentives. Our main result proves that this logic applies more generally as we locally vary the precision of the central bank's private information, away from the extremes of perfect and no information.

[7] Formally, punishment corresponds to "money burning" in principal-agent models of delegation (e.g., Amador and Bagwell, 2013). The agent in our setting is the central bank, and the principal is society, or equivalently the central bank's ex-ante "self" which is not inflation-biased.

[8] See for example Innes (1990) and Levin (2003). Here these incentives arise because punishments cannot depend directly on the central bank's policy under a target-based rule.

We additionally show that the benefit of using a target-based rule over an instrument-based rule is decreasing in the severity of the central bank's commitment problem, namely decreasing in its inflation bias and increasing in the severity of punishment. Intuitively, the less biased is the central bank, the less costly is incentive provision under a target-based rule, as the central bank can be deterred from choosing excessively expansionary policy with less frequent punishments. Similarly, the harsher is the punishment imposed on the central bank for exceeding the inflation threshold, the less often are punishments required on path to implement a target outcome. We show that these two forces therefore make target-based rules more appealing than instrument-based rules on the margin.

Our results shed light on the adoption of inflation targeting around the world. In particular, our analysis suggests that inflation targeting may be more suitable to advanced economies where the central bank commitment problem tends to be less severe compared to emerging market economies.[9] Fraga, Goldfajn, and Minella (2003) indeed find that inflation targeting has been comparatively less successful in emerging markets, where deviations from the target have been larger and more common than in advanced economies. At the same time, our findings can also justify the more recent adoption of inflation targeting rules in emerging markets which have experienced an improvement in central bank independence. Viewing the central bank's bias as fueled by political interests, such independence implies a smaller bias and thus greater benefits from adopting inflation targeting.

A natural question from our analysis is how instruments and targets can be combined to improve upon the above rules that rely exclusively on one of these tools. We study the optimal hybrid rule, that is the optimal unconstrained rule in which punishments can depend freely on the central bank's policy and realized inflation.[10] Perhaps surprisingly, we show that this rule admits a simple form: an instrument threshold that is relaxed whenever a target threshold is satisfied. The optimal hybrid rule dominates instrument-based rules by allowing the central bank more flexibility to use expansionary policy under a target-based criterion, and it dominates target-based rules by more efficiently limiting the central bank's discretion with direct punishments. An implementation of the optimal hybrid rule would be a Taylor rule which, whenever violated, switches to an inflation target. Notably, some policymakers and economists

[9]Relatedly, Mishkin (2000) and Taylor (2014) also conclude that inflation targeting may not be appropriate for many emerging-market countries.

[10]This rule yields the highest social welfare that can be achieved given the central bank's inflation bias and private information.

advocated such an approach in the US in the aftermath of the Global Financial Crisis; see for example Yellen (2015, 2017) and the discussion in Walsh (2015).[11]

This paper is related to several literatures. It relates to the literature on central bank credibility and reputation pioneered by Rogoff (1985), including work that examines the role of private information in such a context.[12] We follow Athey, Atkeson, and Kehoe (2005) by taking a mechanism design approach to characterize optimal policy subject to private information constraints. The paper also fits into the mechanism design literature that studies the tradeoff between commitment and flexibility in principal-agent delegation settings, building on the seminal work of Holmström (1977, 1984). In particular, in addition to Athey, Atkeson, and Kehoe (2005), the delegation models of Waki, Dennis, and Fujiwara (2018) on monetary policy and Amador, Werning, and Angeletos (2006) and Halac and Yared (2014, 2018, 2019) on fiscal policy are closely related.[13] Our main departure from these two literatures is that we distinguish between incentives that are based on an agent's actions (as in these literatures) and incentives that are based on observable noisy outcomes, and we characterize optimal incentives which can condition on both variables.[14] Additionally, we depart from these literatures in that we allow for limited enforcement of rules, since incentives are provided via a finite punishment in our model.[15]

Prior work has considered the optimal choice of policy instrument, as well as the optimal choice of policy target. For a starting point on instruments, see Friedman (1960) on money growth rules, McCallum (1981) and Taylor (1993) on interest rate rules, and Devereux and Engel (2003) on exchange rate rules. Poole (1970) and Weitzman (1974) provide a broad discussion of instrument rules based on prices—such as interest rates or exchange rates—versus instrument rules based on quantities—such

[11]The policy of the Italian government after it joined the European Exchange Rate Mechanism (ERM) in 1979 also resembled an optimal hybrid rule of this form. In 1992, the Italian government left the ERM to respond to its weak economy and devalue its currency. It subsequently pursued a more expansionary monetary policy that targeted inflation, until it returned to the ERM in 1996. See Bank of Italy (1996), Passacantando (1996), and Buttiglione, Giovane, and Gaiotti (1997).

[12]For example, see Barro and Gordon (1983a,b), Backus and Driffill (1985), Canzoneri (1985), Cukierman and Meltzer (1986), Walsh (1995), Sleet (2001), Albanesi, Chari, and Christiano (2003) and Kocherlakota (2016), among others.

[13]More generally, see Alonso and Matouschek (2008), Amador and Bagwell (2013), and Ambrus and Egorov (2017).

[14]Waki, Dennis, and Fujiwara (2018) provide an inflation-targeting implementation of their optimal interest-rate rule.

[15]This is important to assess target-based rules, which would otherwise achieve the socially optimal outcome with infinite punishments that occur with virtually zero on-path probability (cf. Mirrlees, 1974). Halac and Yared (2019) studies delegation under limited enforcement in a fiscal policy setting.

as money growth. Atkeson, Chari, and Kehoe (2007) show that the optimal choice of monetary policy instrument depends on tightness and transparency. For a study of policy targets, see the general framework of Giannoni and Woodford (2017), as well as an extensive body of work that examines inflation targeting, as noted in fn. 1. In contrast to these literatures which consider either instrument-based or target-based rules, the goal of our paper is to compare the performance of these different rule classes, with a focus on their implications for incentive provision.

A comparison of policy instruments and policy targets appears in the recent work of Angeletos and Sastry (2019), which examines the optimal form of forward guidance. Their focus however is on bounded rationality, namely households making mistakes in reasoning about others' behavior and the equilibrium effects of policy. Instead, our focus is on the policymaker's lack of commitment and how to optimally provide incentives given the resulting bias in preferences.

By studying incentives under both private information and noisy outcomes, our paper also relates to the work of Riordan and Sappington (1988) and the literature that followed it. In the classic context of a regulator and a privately informed firm, the main question of this literature is whether the first best can be achieved by making transfers dependent on both the firm's behavior and ex-post public signals correlated with the firm's type.[16] These models differ from ours in several aspects, most importantly in that incentives are provided via transfers, which are ruled out in our delegation environment.[17] Moreover, while the focus of this literature is on the conditions under which the first best obtains under hybrid schemes, we characterize and compare second-best rules when costly punishments may condition only on behavior, or only on realized outcomes, or on both.

Finally, our paper is related to other work that considers the use of socially costly punishments as incentives like we do, albeit in different environments. See for example Acemoglu and Wolitzky (2011) and Padró i Miquel and Yared (2012).[18]

[16]The results of this literature are related to those of Cremer and McLean (1985) in auction theory.

[17]Some of these papers consider limited liability constraints; see Kessler, Lülfesmann, and Schmitz (2005) and references therein. Other differences with our work include that values are private in these models and, except for Gary-Bobo and Spiegel (2006), public signals are only informational.

[18]Although the settings and analyses differ in many aspects, our paper also relates to some models of career concerns for experts which study how the information a principal has affects reputational incentives. In particular, Prat (2005) distinguishes between information on actions and on outcomes.

2 Model

We present a monetary policy model in which the central bank observes a demand shock forecast and then chooses policy subject to lack of commitment. The forecast is the central bank's private information, so we refer to it as the central bank's *type*. Inflation and output depend on both the choice of monetary policy and the realized demand shock. Society provides incentives to the central bank via punishment which is socially costly. After describing the environment and our solution concept, we show that the model takes the form of a principal-agent delegation problem, with society being the principal and the central bank the agent.

Setup. Consider a linearized closed economy New Keynesian model with demand shocks (e.g. Benigno and Woodford, 2005). The economy consists of society and a central bank, and the horizon is infinite with time periods $t = 0, 1, \ldots$ The New Keynesian Phillips curve in period t, linking output to inflation, can be written as

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

where $\pi_t \in \mathbb{R}$ is the inflation rate, $x_t \in \mathbb{R}$ is the output gap, $\beta \in (0,1)$ is the social discount factor, and $\kappa > 0$ is the slope of the Phillips curve. The Euler equation, linking output growth rates to real interest rates, is

$$x_t = \mathbb{E}_t x_{t+1} - \zeta(i_t - \mathbb{E}_t \pi_{t+1}) - \theta_t/\kappa, \tag{2}$$

where $i_t \in \mathbb{R}$ is the nominal interest rate, $\zeta > 0$ is the elasticity of intertemporal substitution, and $\theta_t \in \mathbb{R}$ is a zero-mean aggregate demand shock which we normalize by κ to simplify the notation. We assume that θ_t is independent and identically distributed (i.i.d.) over time; we show in Subsection 6.2 that our results extend to a setting in which θ_t is persistent.

In every period t, the central bank privately observes a signal $s_t \in \{s^L, s^H\}$ and chooses the interest rate i_t.[19] The demand shock θ_t is then realized, market expectations of the future are formed, and inflation π_t and the output gap x_t are determined. We remark that while the policy instrument in our closed economy is the interest rate i_t, our results apply directly to an open economy in which the policy instrument is

[19] Taking the signal s_t to be binary allows us to present our main insights in a transparent way. In Section 5, we extend our results to the case in which s_t is drawn from a continuum.

the exchange rate. As discussed in Subsection 6.3, the reason is that the uncovered interest rate parity condition implies a one-to-one mapping between the central bank's choice of exchange rate and its choice of interest rate (Galí, 2015).

The central bank's signal $s_t \in \{s^L, s^H\}$ is informative about the demand shock. Specifically, we assume that the conditional distribution of the shock is normal with mean equal to the signal, i.e., $\theta_t|_{s_t} \sim \mathcal{N}(s_t, \sigma^2)$. The precision of the central bank's information is given by $\sigma^{-1} > 0$. We take $s^L = -\Delta$ and $s^H = \Delta$ for some $\Delta > 0$ and assume that each signal occurs with equal probability. The shock's unconditional distribution is thus a mixture of two normal distributions and has mean and variance given by

$$\mathbb{E}(\theta) = 0 \quad \text{and} \quad Var(\theta) = \sigma^2 + \Delta^2.$$

Society observes the central bank's policy i_t and the realized shock θ_t (or, equivalently, the central bank's policy i_t and the realized inflation π_t). Society cannot however deduce the central bank's type s_t from these observations, as the distribution of θ_t has full support over the entire real line for each signal s_t.

Society can incentivize the central bank by imposing penalties which are mutually costly to the central bank and society itself.[20] We model these penalties by letting society commit ex ante to punishments $P_t(\cdot)$, specifying a cost to be imposed at the end of period t as a function of the central bank's policy and/or realized inflation. As described below, our analysis will focus on Markov equilibria, so we require the punishment function $P_t(\cdot)$ to condition only on period-t variables, namely policy i_t and/or inflation π_t. Additionally, we require $P_t \in [0, \overline{P}]$ for some finite $\overline{P}$.[21] As discussed in the Introduction, the punishments specified by society may represent formal penalties due to sanction regimes as well as informal penalties in the form of a loss of reputation for the central bank. Note that our assumption that the signal s_t is privately observed by the central bank can be equivalently interpreted as a restriction on the punishment function $P_t(\cdot)$, which cannot explicitly condition on the signal.

[20] Our model abstracts from any transfers between society and the central bank, which are rarely observed in practice and are also ruled out in the literature on principal-agent delegation. The punishment in our model being common to society and the central bank captures, in a stark way, the fact that providing the central bank with incentives is socially costly. In Section 6, we discuss an extension in which punishments harm society and the central bank asymmetrically.

[21] Observe that since punishments are socially costly, society's rule will not be renegotiation proof, and thus commitment to $P_t(\cdot)$ is important. This commitment is ensured in practice via procedural rules designed to follow through on sanctions. Our analysis and results could be extended to allow for a small probability of renegotiation, with its effect being analogous to a reduction in $\overline{P}$.

Payoffs and First Best. Expected welfare at date t is given by

$$\mathbb{E}_{t-1}\sum_{k=0}^{\infty}\beta^k\left[\alpha\kappa x_{t+k}-\gamma\frac{(\kappa x_{t+k})^2}{2}-(1-\gamma)\frac{\pi_{t+k}^2}{2}-P_{t+k}\right],$$

for some $\alpha>0, \gamma\in[0,1]$. Substituting with the Phillips curve in (1), this can be rewritten in terms of the sequence of inflation:[22]

$$\mathbb{E}_{t-1}\left\{\alpha\pi_t+\sum_{k=0}^{\infty}\beta^k\left[-\gamma\frac{(\pi_{t+k}-\beta\mathbb{E}_{t+k}\pi_{t+k+1})^2}{2}-(1-\gamma)\frac{\pi_{t+k}^2}{2}\right]-P_{t+k}\right\}. \tag{3}$$

Note that these preferences are time-inconsistent: for $k\geq 1$, π_{t+k} enters differently into welfare from the perspective of date t relative to the perspective of date $t+k$. We follow Benigno and Woodford (2005) by assuming that society cares about welfare from the "timeless perspective." That is, social welfare is given by

$$\mathbb{E}_{-1}\sum_{t=0}^{\infty}\left\{\beta^t\left[-\gamma\frac{(\pi_t-\beta\mathbb{E}_t\pi_{t+1})^2}{2}-(1-\gamma)\frac{\pi_t^2}{2}\right]-P_t\right\}. \tag{4}$$

Instead, the central bank at date t wishes to maximize the welfare function in (3), reflecting the fact that the central bank lacks commitment to policy.

To illustrate the commitment problem, consider the first-best policy that maximizes social welfare in (4). This policy prescribes no punishment, i.e. $P_t=0$ for all $t=0,1,\ldots$, and it implies a "divine coincidence" (e.g. Blanchard and Galí, 2007) as social welfare is maximized with inflation stabilization. Plainly, under full commitment and perfect forecasting of the demand shocks θ_t, setting an interest rate $i_t=-\theta_t/(\kappa\zeta)$ in each period t effectively counteracts the demand shocks, targeting zero inflation and a zero output gap at all dates. In fact, analogous logic applies if the central bank observes only an imperfect signal s_t of the shock θ_t but can commit to its policy choice. In this case, the central bank would set $i_t=-s_t/(\kappa\zeta)$ in each period t in order to guarantee zero expected inflation and a zero expected output gap at all dates.

Things however change when the central bank lacks commitment to policy. In this case, the central bank at date t maximizes its welfare represented by equation (3), which puts additional weight $\alpha>0$ on current inflation π_t compared to the social welfare function in (4). The reason is that, lacking commitment, the central bank at

[22]This representation is derived under the condition that $\lim_{T\to\infty}\mathbb{E}_{t-1}\beta^T\pi_T=0$. As we will describe, we focus on Markov perfect equilibria where this condition holds.

t does not internalize the impact of current inflation on past inflation expectations. As a result, the central bank's policy at t may differ from the socially optimal policy $i_t = -s_t/(\kappa\zeta)$. We will refer to α as the central bank's *inflation bias*.

Equilibrium and Reformulation. Consider policy at date t under lack of commitment by the central bank, given a punishment specified by society as a function of the central bank's policy choice i_t and/or realized inflation π_t. In principle, the central bank's equilibrium strategy could depend on the entire history of signals, demand shocks, inflation, output gaps, interest rates, and punishments, potentially leading to a complicated path of private sector expectations. To make our analysis tractable, we focus on Markov perfect equilibria in which the central bank's strategy at date t depends only on its current signal s_t and the punishment function $P_t(\cdot)$, and in which private sector expectations of the future are constant, as is natural by the signals and shocks being i.i.d. over time.[23]

Under this solution concept, equilibrium policies and payoffs are stationary, and incentives are provided to the central bank via within-period punishments only.[24] To simplify the notation, we thus remove time subscripts hereafter, and we reformulate the problem as follows. Observe that given constant expectations of future inflation $\mathbb{E}\pi$ and output gap $\mathbb{E}x$ (both on- and off-the-equilibrium path), a central bank's observable choice of interest rate i is equivalent to an observable choice of policy μ defined as

$$\mu \equiv (\beta + \kappa\zeta)\mathbb{E}\pi + \kappa\mathbb{E}x - \kappa\zeta i.$$

Moreover, substituting into (1) and (2), inflation is then given by $\pi = \mu - \theta$, expected inflation by $\mathbb{E}\pi = \mathbb{E}\mu$, and the output gap by $x = (\mu - \theta - \beta\mathbb{E}\mu)/\kappa$. Using this notation, we can then redefine the punishment specified by society as a function of μ and $\pi = \mu - \theta$, namely a function $P(\mu, \mu - \theta)$.

Let $\phi(z|\overline{z}, \sigma_z^2)$ be the normal density of a variable z with mean $\overline{z}$ and variance σ_z^2, and let $\Phi(z|\overline{z}, \sigma_z^2)$ be the corresponding normal cumulative distribution function. With the notation just introduced, the per-period social welfare corresponding to (4) can be

[23] We therefore ignore issues relating to potentially unstable expectations and implementation (see Woodford, 2003; Cochrane, 2011; Galí, 2015).

[24] That is, we rule out incentives via self-enforcing continuation play. Halac and Yared (2020) studies optimal equilibria in a monetary model under history-dependent strategies, showing that incentive provision can require the economy to transition between low inflation and high inflation regimes over time.

written as

$$\sum_{j=L,H} \frac{1}{2} \int_{-\infty}^{\infty} \left[U(\mu^j, \theta, \mathbb{E}\mu) - P(\mu^j, \mu^j - \theta) \right] \phi(\theta | s^j, \sigma^2) d\theta, \tag{5}$$

where

$$U(\mu^j, \theta, \mathbb{E}\mu) = -\gamma \frac{\left(\mu^j - \theta - \beta \mathbb{E}\mu\right)^2}{2} - (1-\gamma) \frac{\left(\mu^j - \theta\right)^2}{2}.$$

Recall that the central bank's welfare differs from society's due to the time inconsistency problem. After observing its type s^j, and taking as given the punishment function $P(\mu^j, \mu^j - \theta)$ and expectations $\mathbb{E}\mu$, the central bank chooses a policy μ^j to maximize

$$\int_{-\infty}^{\infty} \left[\alpha \left(\mu^j - \theta \right) + U(\mu^j, \theta, \mathbb{E}\mu) - P(\mu^j, \mu^j - \theta) \right] \phi(\theta | s^j, \sigma^2) d\theta \tag{6}$$

for $j = L, H$.

With this reformulation, it is easy to view our setting as one of principal-agent delegation. Society is the principal and the central bank is the agent. Society has a preferred, or *first-best*, policy which is given by $\mu^{fb}(s) = s$, implying zero expected inflation and a zero expected output gap, i.e., $\mathbb{E}\mu = \mathbb{E}\pi = 0$. The central bank is better informed than society and has a preferred, or *flexible*, policy which is given by $\mu^f(s, \mathbb{E}\mu) = s + \alpha + \gamma\beta\mathbb{E}\mu$.[25] Thus, if granted full flexibility, the central bank's policy would imply strictly positive expected inflation and a strictly positive expected output gap, i.e., $\mathbb{E}\mu = \mathbb{E}\pi > 0$. It is in this sense that the central bank is inflation-biased.

We make two assumptions on parameters. First, we assume that Δ is relatively small, so that the central bank types are relatively "close" to each other:

Assumption 1. *$\alpha \geq 2\Delta$.*

We make this assumption to ensure that our results do not rely on a discrete distance between the central bank types, as we verify in our extension to a continuum of types in Section 5. The implication of Assumption 1 is that, holding $\mathbb{E}\mu = 0$, the central bank's flexible policy exceeds the first-best policy under each signal:

$$s^L < s^H \leq s^L + \alpha < s^H + \alpha. \tag{7}$$

[25]Observe that if $\gamma = 0$, then our setting corresponds to a delegation setting with quadratic preferences and a constant bias (e.g., Melumad and Shibano, 1991; Alonso and Matouschek, 2008).

Second, we assume that society has a sufficient breadth of incentives to use in disciplining the central bank:

Assumption 2. $\overline{P} \geq \frac{1}{2\phi(1|0,1)}\left(\frac{\alpha}{1-\gamma\beta}\right)^2$.

Classes of Rules. We distinguish between different classes of rules according to how incentives are structured. We say that a rule is instrument-based if society commits to a punishment which depends only on the central bank's policy choice μ, that is, a punishment function $P(\mu)$. A rule instead is target-based if society commits to a punishment that depends only on the realized inflation $\pi = \mu - \theta$, that is, a punishment function $P(\mu - \theta)$. Finally, if society commits to a punishment function $P(\mu, \mu - \theta)$ that depends freely on μ and θ (and therefore freely on μ and π), we say that the rule is hybrid.

We are interested in comparing the performance of these different classes of rules as the environment changes. Our analysis will consider varying the precision of the central bank's private information while holding fixed the mean and variance of the shock θ. At one extreme, we can take $\sigma \to \sqrt{Var(\theta)}$ and $\Delta \to 0$, so the central bank is uninformed with signal $s^L = s^H = 0$. At the other extreme, we can take $\sigma \to 0$ and $\Delta \to \sqrt{Var(\theta)}$, so the central bank is perfectly informed with signal $s^j = \theta$.[26] Note that since Assumption 1 holds for all feasible $\sigma > 0$ and $\Delta > 0$ given $Var(\theta)$ fixed, the assumption implies $\alpha \geq 2\sqrt{Var(\theta)} \geq 2\sigma$.

3 Instrument-Based and Target-Based Rules

We examine rules based on instruments versus targets. Subsection 3.1 and Subsection 3.2 solve for the optimal rule within each class. Subsection 3.3 offers a comparison and shows that the socially optimal class of rule depends on the precision of the central bank's private information, the central bank's inflation bias, and the severity of punishment.

[26] Note that in our setting, welfare under an instrument-based or a target-based rule depends only on the mean and variance of θ. This avoids additional complications stemming from the fact that higher moments of the distribution of θ vary with σ and Δ.

3.1 Optimal Instrument-Based Rule

An instrument-based rule specifies a policy μ^j for each central bank type $j = L, H$ and a punishment $P(\mu)$ as a function of the policy choice μ. Let $P^j \equiv P(\mu^j)$ for $j = L, H$. The allocation $\{\mu^L, \mu^H, P^L, P^H\}$ must satisfy private information, enforcement, feasibility, and rationality constraints, as we describe next.

The private information constraint captures the fact that the central bank can misrepresent its type. The rule must be such that, for $j = L, H$, a central bank of type j has no incentive to deviate privately from policy μ^j to policy μ^{-j}:

$$\int_{-\infty}^{\infty} \left[\alpha\mu^j + U(\mu^j, \theta, \mathbb{E}\mu) - P^j\right] \phi(\theta|s^j, \sigma^2)d\theta \\ \geq \int_{-\infty}^{\infty} \left[\alpha\mu^{-j} + U(\mu^{-j}, \theta, \mathbb{E}\mu) - P^{-j}\right] \phi(\theta|s^j, \sigma^2)d\theta. \tag{8}$$

The enforcement constraint captures the fact that the central bank can freely choose any policy $\widetilde{\mu} \in \mathbb{R}$, including policies not assigned to either type. The rule must be such that, for $j = L, H$, a central bank of type j has no incentive to deviate publicly from policy μ^j to any policy $\widetilde{\mu} \notin \{\mu^L, \mu^H\}$:

$$\int_{-\infty}^{\infty} \left[\alpha\mu^j + U(\mu^j, \theta, \mathbb{E}\mu) - P^j\right] \phi(\theta|s^j, \sigma^2)d\theta \\ \geq \int_{-\infty}^{\infty} \left[\alpha\widetilde{\mu} + U(\widetilde{\mu}, \theta, \mathbb{E}\mu) - P(\widetilde{\mu})\right] \phi(\theta|s^j, \sigma^2)d\theta.$$

Note that since the punishment satisfies $P(\mu) \leq \overline{P}$ for all $\mu \in \mathbb{R}$, the above inequality must hold under maximal punishment, i.e. when $P(\mu) = \overline{P}$. Moreover, since the inequality must then hold for all $\widetilde{\mu} \in \mathbb{R}$, it must necessarily hold when $\widetilde{\mu}$ corresponds to type j's flexible policy $\mu^f(s^j, \mathbb{E}\mu) = s^j + \alpha + \gamma\beta\mathbb{E}\mu$. A necessary condition for the enforcement constraint to be satisfied is thus

$$\int_{-\infty}^{\infty} \left[\alpha\mu^j + U(\mu^j, \theta, \mathbb{E}\mu) - P^j\right] \phi(\theta|s^j, \sigma^2)d\theta \\ \geq \int_{-\infty}^{\infty} \left[\alpha\mu^f(s^j, \mathbb{E}\mu) + U(\mu^f(s^j, \mathbb{E}\mu), \theta, \mathbb{E}\mu) - \overline{P}\right] \phi(\theta|s^j, \sigma^2)d\theta \tag{9}$$

for $j = L, H$, where note that the right-hand side is the central bank's minmax payoff.

Constraints (8) and (9) are clearly necessary for an instrument-based rule prescribing $\{\mu^L, \mu^H, P^L, P^H\}$ to be incentive compatible. Furthermore, if these constraints are satisfied, then this allocation can be supported by specifying the worst punishment $\overline{P}$ following any policy choice $\mu \notin \{\mu^L, \mu^H\}$. Since such a choice is *off path*, it is without loss to assume that it is maximally punished.

Finally, feasibility of the rule requires that for $j = L, H$,

$$P^j \in [0, \overline{P}], \tag{10}$$

and rationality of private sector expectations requires

$$\mathbb{E}\mu = \frac{1}{2}\mu^L + \frac{1}{2}\mu^H. \tag{11}$$

An optimal instrument-based rule maximizes expected social welfare subject to the private information, enforcement, feasibility, and rationality constraints:

$$\max_{\mu^L, \mu^H, P^L, P^H} \sum_{j=L,H} \frac{1}{2} \int_{-\infty}^{\infty} \left[U(\mu^j, \theta, \mathbb{E}\mu) - P^j \right] \phi(\theta | s^j, \sigma^2) d\theta \tag{12}$$

$$\text{subject to, for } j = L, H, \text{ (8), (9), (10), and (11)}.$$

Define a *maximally-enforced instrument threshold* μ^* as a rule that prescribes no punishment if the central bank's policy is weakly below a threshold μ^* and the maximal punishment $\overline{P}$ if the policy exceeds this threshold. Such a rule can take the form of a maximally-enforced lower bound on the nominal interest rate in our environment.[27] We find:

Proposition 1. *The optimal instrument-based rule specifies $\mu^L = \mu^H = 0$ and $P^L = P^H = 0$. This rule can be implemented with a maximally-enforced instrument threshold $\mu^* = 0$.*

The optimal instrument-based rule assigns both central bank types the policy that maximizes ex-ante social welfare. The central bank is given no discretion, and punishments occur only off path, if the central bank were to publicly deviate to a different policy. The simple structure of this rule will allow us to compare its performance to that of the optimal target-based rule in a transparent way. Our results however do not rely on this simple structure, as we show in Section 5.[28]

[27] In Subsection 6.2, we show that such a lower bound will be a function of the state in a setting in which demand shocks are persistent over time.

[28] In the setting of Section 5 with a continuum of signals, the optimal instrument-based rule does

It is worth noting that an instrument-based rule that induces the central bank to choose the first-best policy, $\mu^{fb}(s^j) = s^j$ for $j = L, H$, is available to society. Specifically, the low type can be dissuaded from choosing s^H by specifying an on-path punishment $P^H > 0$ (and both types can be dissuaded from choosing any policy $\mu \notin \{s^L, s^H\}$ by specifying the worst punishment off path). However, because punishment is socially costly, Proposition 1 shows that such a rule is strictly dominated by a maximally-enforced instrument threshold $\mu^* = 0$.

To prove Proposition 1, we solve a relaxed version of the program in (12) which ignores the private information constraint (8) for the high type and the enforcement constraints (9) for both types. We show that under Assumption 1, the solution to this relaxed problem entails no discretion, and it thus satisfies (8) for both types. Moreover, Assumption 2 guarantees that (9) is also satisfied for both types.

Proposition 1 relates to the findings of an extensive literature on delegation, which provides conditions under which threshold delegation with no money burning is optimal. A general treatment can be found in Amador and Bagwell (2013). The analysis in Halac and Yared (2019) is also related in that it incorporates enforcement constraints like those in (9) into a delegation framework, and it shows the optimality of maximally-enforced thresholds where on- and off-path violations lead to the worst punishment. In the current setting with binary signals, enforcement constraints are non-binding by Assumption 2, so punishments occur only off path. We address the issues that arise when enforcement constraints bind and punishments occur on path in our extension to a continuum of signals in Section 5.

3.2 Optimal Target-Based Rule

A target-based rule specifies a policy μ^j for each central bank type $j = L, H$ and a punishment $P(\pi)$ as a function of inflation $\pi = \mu - \theta$. Note that $P(\pi)$ is defined for $\pi \in \mathbb{R}$ since θ is normally distributed. The allocation $\{\mu^L, \mu^H, \{P(\pi)\}_{\pi \in \mathbb{R}}\}$ must satisfy incentive compatibility, feasibility, and rationality constraints, as we describe next.

Incentive compatibility requires that the policy prescribed for each central bank type solve this type's welfare-maximization problem. Given its private information and the punishment function specified by society, the central bank takes into account how

provide discretion to the central bank and does involve punishments on path, yet our main insights continue to apply.

its policy affects the distribution of inflation and, thus, punishments. For $j = L, H$, μ^j must satisfy:

$$\mu^j \in \arg\max_{\widetilde{\mu}} \left\{ \int_{-\infty}^{\infty} \left[\alpha\widetilde{\mu} + U(\widetilde{\mu}, \theta, \mathbb{E}\mu) - P(\widetilde{\mu} - \theta)\right] \phi\left(\theta | s^j, \sigma^2\right) d\theta \right\}. \tag{13}$$

Additionally, feasibility of the rule requires

$$P(\pi) \in \left[0, \overline{P}\right] \text{ for all } \pi \in \mathbb{R}, \tag{14}$$

and rationality of private sector expectations requires

$$\mathbb{E}\mu = \frac{1}{2}\mu^L + \frac{1}{2}\mu^H. \tag{15}$$

An optimal target-based rule maximizes expected social welfare subject to the incentive compatibility, feasibility, and rationality constraints:

$$\max_{\mu^L, \mu^H, \{P(\pi)\}_{\pi \in \mathbb{R}}} \sum_{j=L,H} \frac{1}{2} \int_{-\infty}^{\infty} \left[U(\mu^j, \theta, \mathbb{E}\mu) - P(\mu^j - \theta)\right] \phi\left(\theta | s^j, \sigma^2\right) d\theta \tag{16}$$

$$\text{subject to, for } j = L, H, \text{ (13) and (14), and (15)}.$$

Note that integration by substitution yields

$$\int_{-\infty}^{\infty} P(\mu - \theta)\phi\left(\theta | s^j, \sigma^2\right) d\theta = \int_{-\infty}^{\infty} P(\pi)\phi\left(\mu - s^j - \pi | 0, \sigma^2\right) d\pi, \tag{17}$$

where we have used the fact that $\phi\left(\theta | s, \sigma^2\right) = \phi\left(\theta - s | 0, \sigma^2\right)$ since $\phi\left(\cdot\right)$ is the density of a normal distribution. Using (17) to substitute in (13), the first-order condition of the central bank's problem is

$$\alpha - \left(\mu^j - s^j\right) + \gamma\beta\mathbb{E}\mu + \int_{-\infty}^{\infty} P(\pi)\phi'\left(\mu^j - s^j - \pi | 0, \sigma^2\right) d\pi = 0 \text{ for } j = L, H. \tag{18}$$

Condition (18) is necessary for the rule to be incentive compatible. Its solution is $\mu^j = s^j + \delta$ for $j = L, H$ and some $\delta \gtreqless 0$, where δ is independent of the central bank's type j. The latter observation allows us to simplify the problem in (16) as social welfare then also becomes independent of the central bank's type j.

Define a *maximally-enforced target threshold* π^* as a rule that prescribes no punishment if inflation is weakly below a threshold π^* and the maximal punishment $\overline{P}$ if inflation exceeds this threshold. We find:

Proposition 2. *The optimal target-based rule specifies* $\mu^j = s^j + \delta$, $P(\pi) = 0$ *if* $\pi \leq \pi^*$, *and* $P(\pi) = \overline{P}$ *if* $\pi > \pi^*$, *for* $j = L, H$, *some* $\delta \in \left(0, \frac{\alpha}{1-\gamma\beta}\right)$, *and* $\pi^* = \delta + \frac{\sigma^2(1-\gamma\beta)}{\delta[1-\gamma\beta(2-\beta)]}$. *This rule can be implemented with a maximally-enforced target threshold* π^*.

The optimal target-based rule provides incentives to the central bank with a maximally-enforced target threshold π^*. Since a more expansionary monetary policy μ results in higher inflation π in expectation, a central bank of type j responds to this threshold by choosing a policy $s^j + \delta$ which is below its flexible policy $\mu^f(s^j, \mathbb{E}\mu) = s^j + \alpha + \gamma\beta\mathbb{E}\mu$. Note that punishment occurs along the equilibrium path whenever $\pi > \pi^*$, so as to appropriately incentivize the central bank.

Since punishment is socially costly, the optimal target-based rule limits its frequency by keeping the central bank's policy above the first-best level. That is, while a rule that induces the socially optimal policy with $\delta = 0$ is available, Proposition 2 shows that this rule is strictly dominated by one that allows distortions with $\delta > 0$. The proposition also shows that the induced expected inflation is below the threshold, i.e., $\mathbb{E}(\pi) = \delta < \pi^*$. A rule that yields $\mathbb{E}(\pi) = \delta = \pi^*$ would be suboptimal, as it would entail punishing the central bank half of the time (the frequency with which π would exceed π^*). In the optimal rule, realized inflation π exceeds π^* less than half of the time so that punishment occurs less often.

To prove Proposition 2, we follow a first-order approach and solve a relaxed version of the program in (16) that replaces the incentive compatibility constraint (13) with the first-order condition (18) of the central bank's problem. Specifically, we consider a doubly-relaxed problem that takes (18) as a weak inequality constraint (cf. Rogerson, 1985) in order to establish the sign of the Lagrange multiplier on (18) and characterize the solution. We prove that the solution to the relaxed problem takes the threshold form described above, and we show that Assumption 1 and Assumption 2 are sufficient to guarantee the validity of this first-order approach.

High-powered incentives of the form described in Proposition 2 arise in moral hazard settings where, as in our model, rewards and punishments are bounded and enter welfare linearly; see for example Innes (1990) and Levin (2003). These incentives arise here because punishment cannot directly depend on the central bank's policy under a target-based rule. In Section 5, we show that Proposition 2 remains valid when the central bank's signal is not binary but drawn from a continuum: since the central bank's first-order condition implies that social welfare is independent of the central bank's type, social welfare is also independent of the number of central bank types.

3.3 Optimal Class of Rule

Our main result uses the characterizations in Proposition 1 and Proposition 2 to compare the performance of instrument-based and target-based rules. We find that which class of rule is optimal for society depends on the precision of the central bank's private information:

Proposition 3. *Take instrument-based and target-based rules and consider changing σ while keeping $Var(\theta)$ unchanged. There exists $\sigma^* > 0$ such that a target-based rule is strictly optimal if $\sigma < \sigma^*$ and an instrument-based rule is strictly optimal if $\sigma > \sigma^*$. The cutoff σ^* is decreasing in the central bank's bias α and increasing in the worst punishment $\overline{P}$.*

To see the logic, consider how social welfare under each class of rule changes as we vary the precision of the central bank's information σ^{-1}, while keeping the shock variance $Var(\theta)$ unchanged. Since the optimal instrument-based rule gives no flexibility to the central bank to use its private information, social welfare under this rule is invariant to σ. Specifically, by Proposition 1 and $Var(\theta) = \mathbb{E}(\theta^2)$ (since $\mathbb{E}(\theta) = 0$), social welfare under the optimal instrument-based rule is given by

$$-\frac{Var(\theta)}{2},$$

independent of σ.[29] In contrast, using Proposition 2, we can verify that social welfare under the optimal target-based rule is decreasing in σ, that is increasing in the precision of the central bank's information. Intuitively, a better informed central bank can more closely tailor its policy to the shock, and is less likely to trigger punishment by overshooting the inflation threshold specified by society. As a result, a higher precision reduces the volatility of inflation and the social cost of providing high-powered incentives under a target-based rule.

These comparative statics imply that to prove the first part of Proposition 3, it suffices to show that a target-based rule is optimal for high enough precision of the central bank's information whereas an instrument-based rule is optimal otherwise.

[29] This independence will not extend to the setting of Section 5 with a continuum of signals, yet our main results will continue to apply.

Consider the extreme in which the central bank is perfectly informed, that is, $\sigma \to 0$ and $\Delta \to \sqrt{Var(\theta)}$. In this case, the optimal target-based rule sets a threshold $\pi^* = 0$, providing steep incentives and inducing the first-best policy. Note that this rule involves no punishments along the equilibrium path, as a perfectly informed central bank of type $j = L, H$ chooses $\mu^j = \mu^{fb}(s^j) = s^j$ to avoid punishment. Consequently, in this limit case, the optimal target-based rule yields social welfare

$$0 > -\frac{Var(\theta)}{2},$$

and thus it dominates the optimal instrument-based rule.

Consider next the extreme in which the central bank is uninformed, that is, $\sigma \to \sqrt{Var(\theta)}$ and $\Delta \to 0$. In this case, the optimal instrument-based rule guarantees the socially optimal policy given no information by tying the hands of the central bank. Instead, a target-based rule gives the central bank unnecessary discretion and requires punishments to provide incentives. The optimal target-based rule in this limit case sets a threshold $\pi^* > 0$, inducing a central bank of type $j = L, H$ to choose $\mu^j = s^j + \delta$ for $\delta > 0$ and yielding social welfare

$$-\frac{Var(\theta)}{2} - \frac{\delta^2}{2} - \Phi\left(\delta - \pi^* | 0, \sigma^2\right) \overline{P} < -\frac{Var(\theta)}{2}.$$

Thus, this rule is dominated by the optimal instrument-based rule.

The second part of Proposition 3 shows that the benefit of using a target-based rule over an instrument-based rule is decreasing in the central bank's inflation bias and increasing in the severity of punishment. The less biased is the central bank, the less costly is incentive provision under a target-based rule, as relatively infrequent punishments become sufficient to deter excessively expansionary policy. Similarly, the harsher is the punishment experienced by the central bank for missing the target threshold, the less often punishment needs to be used on the equilibrium path to provide incentives under a target-based rule. In contrast, the optimal instrument-based rule is independent of the central bank's bias and the severity of punishment. As such, target-based rules dominate instrument-based rules for a larger range of parameters when the central bank's bias is relatively small or punishment is relatively severe.[30]

[30] One may also wonder how the comparison of the two rule classes changes if the unconditional variance of the shock increases. The answer is ambiguous in that it depends on the extent to which the increase in $Var(\theta)$ is due to an increase in σ or Δ: if only σ increases, instrument-based rules

In addition to the results in Proposition 3, the comparison of instrument-based and target-based rules reveals differences with regards to volatility, specifically to how the two rule classes trade off different kinds of volatility. Relative to the optimal target-based rule, the optimal instrument-based rule induces a lower variance of the central bank's policy, while yielding relatively more volatile output and inflation. At the same time, the optimal instrument-based rule yields lower average inflation than the optimal target-based rule.

4 Optimal Unconstrained Rule

A hybrid rule combines features of instrument-based and target-based rules, with a punishment $P(\mu, \mu-\theta)$ that depends freely on the central bank's policy choice μ and inflation $\mu-\theta$. For $j = L, H$, denote by $P^j(\theta)$ the punishment assigned to central bank type j as a function of the shock θ. Analogous to the program in (12), an optimal hybrid rule solves:

$$\max_{\mu^L,\mu^H,\{P^L(\theta),P^H(\theta)\}_{\theta\in\mathbb{R}}} \sum_{j=L,H} \frac{1}{2}\int_{-\infty}^{\infty} \left[U(\mu^j,\theta,\mathbb{E}\mu) - P^j(\theta)\right]\phi(\theta|s^j,\sigma^2)d\theta \tag{19}$$

subject to, for $j = L, H$,

$$\int_{-\infty}^{\infty} \left[\alpha\mu^j + U(\mu^j,\theta,\mathbb{E}\mu) - P^j(\theta)\right]\phi(\theta|s^j,\sigma^2)d\theta \geq \int_{-\infty}^{\infty} \left[\alpha\mu^{-j} + U(\mu^{-j},\theta,\mathbb{E}\mu) - P^{-j}(\theta)\right]\phi(\theta|s^j,\sigma^2)d\theta, \tag{20}$$

$$\int_{-\infty}^{\infty} \left[\alpha\mu^j + U(\mu^j,\theta,\mathbb{E}\mu) - P^j(\theta)\right]\phi(\theta|s^j,\sigma^2)d\theta \geq \int_{-\infty}^{\infty} \left[\alpha\mu^f(s^j,\mathbb{E}\mu) + U(\mu^f(s^j,\mathbb{E}\mu),\theta,\mathbb{E}\mu) - \overline{P}\right]\phi(\theta|s^j,\sigma^2)d\theta, \tag{21}$$

$$P^j(\theta) \in [0,\overline{P}] \text{ for all } \theta\in\mathbb{R}, \text{ and} \tag{22}$$

$$\mathbb{E}\mu = \frac{1}{2}\mu^L + \frac{1}{2}\mu^H. \tag{23}$$

become more beneficial over target-based rules (as implied by Proposition 3), whereas if only Δ increases, the opposite is true.

The solution to this program yields the highest social welfare that can be achieved subject to the central bank being inflation-biased and privately informed. Note that constraints (20)-(21) are analogous to (8)-(9) in the program that solves for the optimal instrument-based rule, but they now allow the punishment to depend on the shock θ in addition to the central bank's policy μ^j. Note also that analogous to our analysis of the optimal target-based rule, the punishment $P^j(\theta)$ can be equivalently written as a function of inflation, $P^j(\pi)$. We use this formulation in what follows to ease the interpretation.

Define a *maximally-enforced hybrid threshold* $\{\mu^*, \mu^{**}, \{\pi^*(\mu)\}_{\mu\in\mathbb{R}}\}$ as a rule that specifies no punishment if inflation is weakly below a threshold $\pi^*(\mu)$ and the maximal punishment $\overline{P}$ if inflation exceeds $\pi^*(\mu)$, where this threshold depends on the central bank's policy μ and satisfies

$$\pi^*(\mu) = \begin{cases} \infty & \text{if } \mu \leq \mu^* \\ h(\mu) & \text{if } \mu \in (\mu^*, \mu^{**}] \\ -\infty & \text{if } \mu > \mu^{**} \end{cases} \tag{24}$$

for cutoffs $\mu^* < \mu^{**}$ and some continuous function $h(\mu) \in (-\infty, \infty)$ with $\lim_{\mu \downarrow \mu^*} h(\mu) = \infty$. The cutoff μ^* is a soft instrument threshold, where any policy $\mu \leq \mu^*$ is not punished independently of inflation. The cutoff $\mu^{**} > \mu^*$ is a hard instrument threshold, where any policy $\mu > \mu^{**}$ is maximally punished independently of inflation. Intermediate policies $\mu \in (\mu^*, \mu^{**}]$ are not punished if inflation satisfies $\pi \leq \pi^*(\mu)$ and maximally punished if inflation satisfies $\pi > \pi^*(\mu)$. Therefore, an interior target threshold only applies to intermediate policy choices.

We find:

Proposition 4. *The optimal hybrid rule specifies $\mu^L < \mu^H$, $P^L(\pi) = 0$ for all π, $P^H(\pi) = 0$ if $\pi \leq \pi^*(\mu^H)$, and $P^H(\pi) = \overline{P}$ if $\pi > \pi^*(\mu^H)$, for some $\pi^*(\mu^H) \in (-\infty, \infty)$. This rule can be implemented with a maximally-enforced hybrid threshold $\{\mu^*, \mu^{**}, \{\pi^*(\mu)\}_{\mu\in\mathbb{R}}\}$, where $\mu^* = \mu^L$ and $\mu^{**} = \mu^H$.*

The optimal hybrid rule prescribes a tight monetary policy and no punishment for the low central bank type, while specifying an expansionary monetary policy and a target threshold for the high central bank type. To prove this result, we solve a relaxed version of (19)-(23) which ignores the private information constraint (20) for the high type and the enforcement constraints (21) for both types. We establish that the solution to this relaxed problem takes the form described in Proposition 4 and satisfies these constraints.

As implied by Proposition 4, the optimal hybrid rule strictly improves upon the rules studied in Section 3. Intuitively, this rule dominates instrument-based rules by giving the central bank more flexibility to respond to its private information while preserving incentives. The reason is that, under a hybrid rule, society can allow the central bank to choose policies $\mu > \mu^*$ and still deter excessively expansionary policies by using a target-based criterion. Analogously, the optimal hybrid rule dominates target-based rules by more efficiently limiting the central bank's discretion to choose policies that are excessively loose. The reason is that, under a hybrid rule, society can avoid punishments under policies $\mu \leq \mu^*$ and directly punish the central bank for policies $\mu > \mu^{**}$, thus reducing the frequency of punishment on path.

In principle, combining instruments and targets could yield rules with complicated forms. Proposition 4 however shows that the optimal hybrid rule admits an intuitive implementation. This rule essentially consists of an instrument threshold μ^* which is relaxed to μ^{**} whenever the target threshold is satisfied. Interestingly, as discussed in the Introduction, monetary policy rules of this form were advocated in the US in the aftermath of the Global Financial Crisis.[31]

5 Continuum of Types

Our baseline model facilitates a transparent analysis by assuming that the central bank's signal of the demand shock is binary. In this section, we show that our main insights remain valid under a continuum of signals.

Suppose that instead of observing a binary signal $s \in \{-\Delta, \Delta\}$ of the demand shock $\theta \in \mathbb{R}$, the central bank observes a signal as rich as the shock: $s \in \mathbb{R}$ with $s \sim \mathcal{N}(0, \Delta^2)$. The shock's conditional distribution obeys $\theta|_s \sim \mathcal{N}(s, \sigma^2)$, and thus the unconditional distribution has $\mathbb{E}(\theta) = 0$ and $Var(\theta) = \sigma^2 + \Delta^2$ as in our baseline model. We define the optimal instrument-based and target-based rules analogously as in Section 3, with formal representations provided in the Online Appendix.

In what follows, we take $\gamma \in \{0, 1\}$. The case of $\gamma = 0$ corresponds exactly to the broadly studied delegation setting with quadratic preferences and a constant bias, as in Melumad and Shibano (1991) and Alonso and Matouschek (2008) among others. The

[31] The optimal hybrid rule also resembles the monetary policy adopted by the Italian government in the 1990s; see fn. 11.

case of $\gamma = 1$ instead brings the model close to a New Keynesian setting with negligible costs of inflation, as suggested by the work of Nakamura et al. (2018). In both cases, we can show that the expectations constraint—namely the analog of constraint (11) under an instrument-based rule and correspondingly constraint (15) under a target-based rule—is no longer binding. This simplifies the analysis and allows us to derive the following characterization.[32]

Proposition 5. *Consider an environment with a continuum of central bank types and $\gamma \in \{0, 1\}$. Then:*

1. *An optimal instrument-based rule specifies $s^*, s^{**} \in (-\infty, \infty)$ such that*

$$\{\mu(s), P(s)\} = \begin{cases} \{\mu^f(s, \mathbb{E}\mu), 0\} & \text{if } s < s^*, \\ \{\mu^f(s^*, \mathbb{E}\mu), 0\} & \text{if } s \in [s^*, s^{**}], \\ \{\mu^f(s, \mathbb{E}\mu), \overline{P}\} & \text{if } s > s^{**}, \end{cases}$$

where

$$\int_{-\infty}^{\infty} \left[\alpha\mu^f(s^*, \mathbb{E}\mu) + U(\mu^f(s^*, \mathbb{E}\mu), \theta, \mathbb{E}\mu)\right] \phi(\theta|s^{**}, \sigma^2) d\theta$$
$$= \int_{-\infty}^{\infty} \left[\alpha\mu^f(s^{**}, \mathbb{E}\mu) + U(\mu^f(s^{**}, \mathbb{E}\mu), \theta, \mathbb{E}\mu) - \overline{P}\right] \phi(\theta|s^{**}, \sigma^2) d\theta.$$

This rule can be implemented with a maximally-enforced instrument threshold $\mu^* = \mu^f(s^*, \mathbb{E}\mu)$.

2. *An optimal target-based rule specifies $\mu(s) = s + \delta$, $P(\pi) = 0$ if $\pi \leq \pi^*$, and $P(\pi) = \overline{P}$ if $\pi > \pi^*$, for some $\delta \in \left(0, \frac{\alpha}{1-\gamma\beta}\right)$ and $\pi^* = \delta + \frac{\sigma^2(1-\gamma\beta)}{\delta[1-\gamma\beta(2-\beta)]}$. This rule can be implemented with a maximally-enforced target threshold π^* which coincides with that in Proposition 2.*

The optimal instrument-based and target-based rules under a continuum of central bank types take the same implementations as under binary types. There are differences, however, with regards to the central bank's behavior under the optimal instrument-based rule. In the case of binary types, Proposition 1 shows that both types of the central bank are bunched at the threshold μ^*, with no discretion and with

[32] We assume that an optimal instrument-based rule is piecewise continuously differentiable. Also, if the program solving for this rule admits multiple solutions that differ only on a countable set of types, we select the solution that maximizes social welfare for those types.

no punishments on path. These features change with a continuum of types. Proposition 5 shows that in this case, central bank types $s \in [s^*, s^{**}]$ are bunched at the threshold μ^* and assigned no punishment. But central bank types $s < s^*$ are given discretion: they satisfy the instrument threshold strictly by choosing their flexible policy $\mu^f(s, \mathbb{E}\mu)$ while enjoying no punishment. Moreover, central bank types $s > s^{**}$ are punished on path: these types break the instrument threshold by choosing their flexible policy $\mu^f(s, \mathbb{E}\mu) > \mu^*$ and are assigned the maximal punishment $\overline{P}$. The cutoff s^{**} corresponds to the type that is indifferent between abiding to the instrument threshold and receiving no punishment versus breaking the threshold and receiving maximal punishment, as shown by the indifference condition in Proposition 5.

Compared to the analysis in Subsection 3.1, characterizing the optimal instrument-based rule under a continuum of types requires more involved arguments. These arguments are related to those developed in Halac and Yared (2019), and we present them in the Online Appendix in three main steps.[33] First, we show that the linearity of social welfare and central bank welfare in the punishment, together with the richness of the information structure, imply that only zero and maximal punishment $\overline{P}$ are prescribed in an optimal instrument-based rule; that is, optimal punishments are *bang-bang*. Second, appealing to the log concavity of the normal density function, we establish that optimal punishments are also monotonic, with central bank types below some cutoff s^{**} receiving zero punishment and central bank types above s^{**} receiving maximal punishment $\overline{P}$.[34] As an implication, types $s > s^{**}$ must choose their flexible policy $\mu^f(s, \mathbb{E}\mu)$. Finally, we show that for $s \leq s^{**}$, the optimal rule prescribes policy that is continuous in the central bank's type and takes the form of a threshold.

The characterization of the optimal target-based rule under a continuum of types follows the same steps as in Subsection 3.2. In fact, Proposition 5 shows that the optimal target threshold π^* is quantitatively identical to that in the case of binary types. As mentioned in Subsection 3.2, the reason is that the first-order condition of the central bank's problem in (18) implies a choice $\delta = \mu - s$ by the central bank which is independent of its type. This means that the optimal target-based rule is independent of the distribution of the central bank's signal s, and therefore the

[33]The arguments in Halac and Yared (2019) cannot be applied directly to the current setting as the agent's bias in that paper takes a multiplicative form. We provide a complete proof in the Online Appendix for the additive bias specification of this paper.

[34]We note that our proof in the Online Appendix, and thus the characterization of the optimal instrument-based rule in Proposition 5, apply more broadly to any distribution of types with a log concave density. This includes the exponential, gamma, log-normal, Pareto, and uniform distributions among others; see Bagnoli and Bergstrom (2005).

characterization in Proposition 2 holds independently of the number of types.

The next proposition uses the characterizations in Proposition 5 to compare the performance of instrument-based and target-based rules. We show that our main result in Proposition 3 continues to apply under a continuum of types: target-based rules dominate instrument-based rules if and only if the central bank's private information is sufficiently precise.

Proposition 6. *Consider an environment with a continuum of central bank types and* $\gamma \in \{0, 1\}$. *Take instrument-based and target-based rules and consider changing* σ *while keeping* $Var(\theta)$ *unchanged. There exists* $\sigma^* > 0$ *such that a target-based rule is strictly optimal if* $\sigma < \sigma^*$ *and an instrument-based rule is strictly optimal if* $\sigma > \sigma^*$.

The result follows from comparing how an increase in the precision of the central bank's signal affects social welfare under the optimal instrument-based rule versus the optimal target-based rule. The argument is more involved than in the case of binary types, as now a more precise signal strictly improves social welfare under both classes of rules. Specifically, in the case of the optimal target-based rule, the effect takes the same form as in Subsection 3.3: if the central bank's information is more precise, then the central bank can better tailor its policy to the shock, which lowers inflation volatility and thus also the need to utilize costly punishments on path. In the case of the optimal instrument-based rule, the effect is richer than in Subsection 3.3: a higher signal precision now allows a better tailoring of the policy to the shock for central bank types $s < s^*$ and $s > s^{**}$ (for cutoffs s^* and s^{**} defined in Proposition 5), therefore lowering inflation volatility under this rule too.

We prove that the social welfare effect of making the central bank's signal more precise is smaller under the optimal instrument-based rule than under the optimal target-based rule. Intuitively, while inflation volatility declines with precision for types $s < s^*$ and $s > s^{**}$ under the instrument-based rule, there are no-target based punishments whose frequency is reduced as a result; moreover, a higher signal precision has no effect on types $s \in [s^*, s^{**}]$ who have no discretion under this rule. In contrast, under the target-based rule, inflation volatility declines with precision for *all* types, and this has the additional effect of reducing target-based punishments on path. Consequently, we are able to show that the relative benefit of using a target-based rule increases with the precision of the central bank's signal. This comparison in turn allows us to obtain the result in Proposition 6 and hence to establish the robustness of our findings to a continuum of central bank types.

6 Extensions

We discuss three extensions of our baseline model. In the first extension, we relax the assumption that punishments are common to the central bank and society by considering asymmetric punishment costs. In the second extension, we study what happens when the shocks affecting the economy are not i.i.d. but persistent over time. Finally, in the third extension, we consider other policy instruments that may be available to the central bank in addition to the interest rate. In all of these, we show that our main findings continue to hold.

6.1 Asymmetric Punishments

Society may be able to specify punishments that harm the central bank more severely than society itself. Consider an extension of our baseline model in which varying the punishment has a larger effect on the central bank's welfare compared to social welfare. Specifically, let us modify the central bank's welfare in (6) so that it is now given by

$$\int_{-\infty}^{\infty} \left[\alpha \left(\mu^j - \theta\right) + U(\mu^j, \theta, \mathbb{E}\mu) - cP(\mu^j, \mu^j - \theta)\right] \phi(\theta | s^j, \sigma^2) d\theta \tag{25}$$

for $j = L, H$ and some $c \in (1, \alpha/2\Delta)$.[35] Note that our baseline model corresponds to setting $c = 1$. Under $c > 1$, any punishment $P(\mu^j, \mu^j - \theta)$ implies a larger cost on the central bank than on society.

This asymmetric-punishments setting yields optimal instrument-based and target-based rules which take the same implementations as under symmetric punishments. The optimal instrument-based rule is in fact quantitatively identical to that in Subsection 3.1, with $\mu^L = \mu^H = 0$ and $P^L = P^H = 0$. Intuitively, given $c < \alpha/2\Delta$, it is still the case that providing incentives to separate the two central bank types is too costly for society, so the optimal instrument-based rule continues to bunch both types. The optimal target-based rule takes the same form as in Subsection 3.2, with high-powered incentives being optimal as the central bank's welfare remains linear in the punishment under (25). Nonetheless, this rule is not quantitatively identical to that under symmetric punishments. The reason is intuitive: since imposing any given punishment on the central bank, and therefore providing incentives, is now less costly for society, the optimal target-based rule induces a policy closer to first best (i.e., with

[35] Our modeling of asymmetric punishments follows Amador and Bagwell (2013) in their study of delegation under imperfect transfers.

a lower value of δ) in the case of asymmetric punishments.

Our main result in Proposition 3 also extends to this asymmetric-punishments setting: target-based rules dominate instrument-based rules if and only if the central bank's private information is sufficiently precise. The argument is the same as in the case of symmetric punishments. In particular, note that since the optimal instrument-based rule provides no discretion to the central bank, social welfare under this rule is invariant to the precision of the central bank's information, as in Subsection 3.3. Moreover, since the optimal target-based rule takes the same form as under symmetric punishments, social welfare under this rule increases with the precision of the central bank's information, also as in Subsection 3.3. Based on these comparative statics, it follows that our main insights are robust to asymmetric punishments.

6.2 Persistent Shocks

We have assumed that the demand shock θ_t is i.i.d. over time. We next show that our results remain valid if the shock is instead persistent. In fact, we find that by suitably reformulating society's problem, our analysis can be applied without change.

We model persistent shocks by letting the realization of the shock at time $t-1$ shift the mean of the shock at time t. That is, suppose $\mathbb{E}[\theta_t|\theta_{t-1}] = f(\theta_{t-1})$ for a weakly increasing function f and $t = 1, 2, \ldots$, and let the signal s_t also be shifted by θ_{t-1} so that $s_t \in \{f(\theta_{t-1}) - \Delta, f(\theta_{t-1}) + \Delta\}$. As in our baseline model, we assume that each signal is realized with equal probability and $\theta_t|_{s_t} \sim \mathcal{N}(s_t, \sigma^2)$.

We focus our attention on Markov perfect equilibria in which the central bank's strategy at date t depends only on s_t and the punishment function $P_t(\cdot)$ specified by society. Furthermore, we consider equilibria within this class in which the distributions of inflation and the output gap in every period are constant (while the distribution of future interest rates may vary over time). These criteria are desirable in that they are satisfied in the first-best allocation, which features zero expected inflation, a zero expected output gap, and an interest rate that co-moves with the signal s_t.

To show that an equilibrium as just defined exists, let $\mathbb{E}\pi$ and $\mathbb{E}x$ denote the time-invariant means of inflation and the output gap respectively. We can define $\widetilde{\mu}_t$ in this environment to be analogous to μ_t in our baseline model, except for the mean of θ_t which is now given by $f(\theta_{t-1})$ instead of zero:

$$\widetilde{\mu}_t \equiv f(\theta_{t-1}) + (\beta + \kappa\zeta)\mathbb{E}\pi + \kappa\mathbb{E}x - \kappa\zeta i_t.$$

Additionally, define $\widetilde{\theta}_t \equiv \theta_t - f(\theta_{t-1})$ and $\widetilde{s}_t \equiv s_t - f(\theta_{t-1})$. Note that the distributions of $\widetilde{\theta}_t$ and $\widetilde{s}_t$ are i.i.d. and the same as in our baseline model. It follows by substitution into (1) and (2) that $\pi_t = \widetilde{\mu}_t - \widetilde{\theta}_t$, $\mathbb{E}\pi_t = \mathbb{E}\widetilde{\mu}_t$, and $x_t = (\widetilde{\mu}_t - \widetilde{\theta}_t - \beta\mathbb{E}\widetilde{\mu}_t)/\kappa$. Therefore, π_t and x_t are guaranteed to be stationary if $\widetilde{\mu}_t$ is stationary, in which case we can ignore time subscripts. The punishment function specified by society is $P(\widetilde{\mu}, \widetilde{\mu} - \widetilde{\theta})$, and per-period social welfare can be represented as

$$\sum_{j=L,H} \frac{1}{2} \int_{-\infty}^{\infty} [U(\widetilde{\mu}^j, \widetilde{\theta}, \mathbb{E}\widetilde{\mu}) - P(\widetilde{\mu}^j, \widetilde{\mu}^j - \widetilde{\theta})] \phi(\widetilde{\theta}|\widetilde{s}^j, \sigma^2) d\theta$$

$$\text{where } U(\widetilde{\mu}^j, \widetilde{\theta}, \mathbb{E}\widetilde{\mu}) = -\gamma \frac{(\widetilde{\mu}^j - \widetilde{\theta} - \beta\mathbb{E}\widetilde{\mu})^2}{2} - (1-\gamma)\frac{(\widetilde{\mu}^j - \widetilde{\theta})^2}{2}.$$

Finally, given a signal s^j and taking as given the punishment function $P(\widetilde{\mu}^j, \widetilde{\mu}^j - \widetilde{\theta})$ and expectations $\mathbb{E}\widetilde{\mu}$, the central bank chooses a policy μ^j to maximize

$$\int_{-\infty}^{\infty} [\alpha(\widetilde{\mu}^j - \widetilde{\theta}) + U(\widetilde{\mu}^j, \widetilde{\theta}, \mathbb{E}\widetilde{\mu}) - P(\widetilde{\mu}^j, \widetilde{\mu}^j - \widetilde{\theta})] \phi(\widetilde{\theta}|\widetilde{s}^j, \sigma^2) d\theta$$

for $j = L, H$. With the variables and objective functions reformulated in this way, it is clear that we can apply the analysis from our baseline model without further change. Consequently, all of our results remain valid in this environment with persistent demand shocks. It is worth noting that, in this environment, the implementation of the optimal instrument-based rule as a maximally-enforced lower bound on the nominal interest rate would entail a bound that varies with the state θ_{t-1}.

6.3 Other Instruments

In our baseline model, the central bank has access to only one policy instrument, namely the interest rate. We next discuss two extensions of our environment in which other policy instruments are available.

First, suppose that instead of studying a closed economy, we consider an open-economy New Keynesian setting (e.g. Woodford, 2003; Galí, 2015). Such a setting is characterized by the Phillips curve in (1), the Euler equation in (2), and an uncovered interest rate parity condition given by

$$i_t = i_t^* + \mathbb{E}_t(\Delta e_{t+1}), \tag{26}$$

where i_t is the domestic interest rate, i_t^* is the interest rate abroad, and Δe_{t+1} is the exchange rate devaluation. Assume that i_t^* is known at date t when the central bank makes its policy choice. Then by (26), a choice of exchange rate devaluation Δe_{t+1} by the central bank implies a choice of domestic interest rate i_t. Since welfare in this economy depends on inflation π_t and the output gap x_t exactly as in our baseline model, it follows that our analysis applies: the optimal instrument-based rule can be implemented as a maximally-enforced lower bound on the domestic interest rate or, equivalently, a maximally-enforced upper bound on exchange rate devaluation.

Second, suppose we consider a closed-economy setting with money growth. Such a setting is characterized by the Phillips curve in (1), the Euler equation in (2), and a money demand equation that can be represented dynamically in terms of money growth Δm_t (e.g., Woodford, 2003; Galí, 2015):

$$\Delta m_t = \pi_t + x_t - \eta i_t - x_{t-1} + \eta i_{t-1}, \tag{27}$$

for $\eta > 0$. Given i_{t-1} and given equations (1) and (2) for π_t and x_t respectively, a choice of money growth Δm_t by the central bank implies a monetary stance that maps into a level of inflation and output gap. Thus, as above, this environment is analogous to our baseline model and our analysis applies: the optimal instrument-based rule puts a cap on how expansionary monetary policy can be, so it can be implemented as a maximally-enforced upper bound on money growth.

7 Concluding Remarks

In this paper, we have studied the optimal design of monetary policy rules. We considered how to provide incentives to a central bank which has non-contractible information about shocks to the economy and cannot commit to policy. Using a mechanism design approach, we characterized optimal instrument-based and target-based rules, examined the conditions under which each rule class is optimal, and solved for the optimal unconstrained rule that combines instruments and targets. Our results imply that inflation targeting should be adopted if the central bank's information is sufficiently precise and its commitment problem is not too severe; otherwise an interest rate rule would perform better. This same conclusion applies in the context of other policy targets, such as output gap or price level targets (which are equivalent to an inflation target in our model), as well as in the context of other policy instruments, such as an exchange rate or a monetary growth rate (as discussed in Subsection 6.3).

We presented a stylized model with binary signals and symmetric punishments, and we showed that our main results extend to settings with a continuum of signals or asymmetric punishments. Our analysis could be further extended in various directions. One extension could explore different specifications of asymmetric punishments by allowing for non-linearities. For example, while society may suffer a lower cost than the central bank from any given punishment, as in the setting discussed in Subsection 6.1, society's cost may also increase at a relatively faster rate with the severity of punishment, becoming closer to the central bank's cost at higher levels. Note that different specifications could give rise to different *penal codes*. In particular, unlike the "bang-bang" punishments that we found to be optimal in our model, the solution could feature punishment that "fits the crime."

Another extension could consider a central bank whose inflation bias is unknown to society, possibly admitting not only different levels but also different signs. Under binary signals and assumptions analogous to Assumption 1 and Assumption 2, an instrument-based rule that bunches the central bank types regardless of their bias would be optimal, so our characterization in Proposition 1 would remain valid. A characterization of the optimal target-based rule, on the other hand, would have to deal with the problem that assigned policies may now depend on the central bank's unknown bias.

While our focus has been on monetary policy—an area where both instruments and targets have been widely used in practice and extensively discussed in the literature—our analysis also holds lessons for other applications.[36] For example, in the context of fiscal policy, rules may place constraints on instruments like spending or on targets like deficits (see National Conference of State Legislatures, 1999). Bohn and Inman (1996) document that a number of US states use rules based on instruments in the form of beginning-of-the-year fiscal requirements, whereas other US states use a target-based criterion in the form of end-of-the-year fiscal requirements. Our analysis suggests that target-based rules will be relatively more favorable if the government has sufficiently precise non-contractible information about future revenues and its deficit bias is not too large. Moreover, a hybrid rule specifying a spending cap that switches to a deficit

[36]The questions we study are relevant for the design of environmental policies. Environmental regulation may focus on technology mandates—requirements on firms' production processes, such as the choice of equipment—or on performance standards—requirements on output, such as maximum emission rates. See, for example, US Congress, Office of Technology Assessment (1995) and Goulder and Parry (2008).

target when violated would dominate both instrument-based and target-based fiscal policy rules. Notably, such a hybrid rule resembles the rule underlying the Swiss debt brake (Beljean and Geier, 2013).

References

Acemoglu, D. and A. Wolitzky (2011): "The Economics of Labor Coercion," *Econometrica*, 79, 555–600.

Albanesi, S., V. V. Chari, and L. J. Christiano (2003): "Expectation Traps and Monetary Policy," *Review of Economic Studies*, 70, 715–741.

Alonso, R. and N. Matouschek (2008): "Optimal Delegation," *Review of Economic Studies*, 75, 259–93.

Amador, M. and K. Bagwell (2013): "The Theory of Delegation with an Application to Tariff Caps," *Econometrica*, 81, 1541–600.

Amador, M., I. Werning, and G.-M. Angeletos (2006): "Commitment Vs. Flexibility," *Econometrica*, 74, 365–396.

Ambrus, A. and G. Egorov (2017): "Delegation and Nonmonetary Incentives," *Journal of Economic Theory*, 171, 101–135.

Angeletos, G.-M. and K. A. Sastry (2019): "Managing Expectations: Instruments vs. Targets," Working Paper.

Appelbaum, B. (2014): "Yellen Says Restraining the Fed's Oversight Would be a 'Grave Mistake'," New York Times, `https://www.nytimes.com/2014/07/17/business/yellen-says-constraining-fed-would-be-a-grave-mistake.html`.

Athey, S., A. Atkeson, and P. J. Kehoe (2005): "The Optimal Degree of Discretion in Monetary Policy," *Econometrica*, 73, 1431–75.

Atkeson, A., V. V. Chari, and P. J. Kehoe (2007): "On the Optimal Choice of a Monetary Policy Instrument," Federal Reserve Bank of Minneapolis Research Department Staff Report 394.

Backus, D. and J. Driffill (1985): "Rational Expectations and Policy Credibility Following a Change in Regime," *Review of Economic Studies*, 52, 211–21.

Bagnoli, M. and T. Bergstrom (2005): "Log-Concave Probability and Its Applications," *Economic Theory*, 26, 445–69.

Bank of Italy (1996): "Ordinary General Meeting of Shareholders, Held in Rome on 31 May 1996," Abridged Report for the Year 1995.

Barro, R. J. and D. B. Gordon (1983a): "A Positive Theory of Monetary Policy in a Natural Rate Model," *Journal of Political Economy*, 91, 589–610.

——— (1983b): "Rules, Discretion and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics*, 12, 101–121.

Beljean, T. and A. Geier (2013): "The Swiss Debt Brake – Has It Been a Success?" *Swiss Journal of Economics and Statistics*, 149, 115–135.

Benigno, P. and M. Woodford (2005): "Inflation Stabilization and Welfare: The Case of a Distorted Steady State," *Journal of the European Economic Association*, 3, 1185–1236.

Bernanke, B. S. (2004): "Remarks by Governor Ben S. Bernanke," Before the National Economists Club, Washington, D.C. December 2, 2004.

——— (2012): "Transcript of Chairman Bernanke's Press Conference," `https://www.federalreserve.gov/mediacenter/files/FOMCpresconf20120125.pdf`.

——— (2015): "The Taylor Rule: A Benchmark for Monetary Policy?" Blog entry: `https://www.brookings.edu/blog/ben-bernanke/2015/04/28/the-taylor-rule-a-benchmark-for-monetary-policy/`.

Bernanke, B. S., T. Laubach, F. S. Mishkin, and A. S. Posen (1999): *Inflation Targeting: Lessons from the International Experience*, Princeton, N.J., Princeton University Press.

Bernanke, B. S. and F. S. Mishkin (1997): "Inflation Targeting: A New Framework for Monetary Policy," *Journal of Economic Perspectives*, 11, 97–116.

Blanchard, O. and J. Galí (2007): "Real Wage Rigidities and the New Keynesian Model," *Journal of Money, Credit and Banking*, 39, 35–65.

Blinder, A. S. (2014): "An Unnecessary Fix for the Fed," The Wall Street Journal, `https://www.wsj.com/articles/alan-s-blinder-an-unnecessary-fix-for-the-fed-1405639582`.

Bohn, H. and R. P. Inman (1996): "Balanced-Budget Rules and Public Deficits: Evidence from U.S. States," *Carnegie-Rochester Conference Series on Public Policy*, 45, 13–76.

Buttiglione, L., P. D. Giovane, and E. Gaiotti (1997): "The Implementation of Monetary Policy in Italy: The Role of Repo Operations and Official Rates," BIS Conference Paper.

Canzoneri, M. B. (1985): "Monetary Policy Games and the Role of Private Information," *American Economic Review*, 75, 1056–1070.

Cochrane, J. H. (2011): "Determinacy and Identification with Taylor Rules," *Journal of Political Economy*, 119, 565–615.

Cremer, J. and R. P. McLean (1985): "Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist When Demands Are Interdependent," *Econometrica*, 53, 345–361.

Cukierman, A. and A. H. Meltzer (1986): "A Theory of Ambiguity, Credibility, and Inflation under Discretion and Asymmetric Information," *Econometrica*, 54, 1099–128.

Devereux, M. B. and C. Engel (2003): "Monetary Policy in the Open Economy Revisited: Price Setting and Exchange-Rate Flexibility," *The Review of Economic Studies*, 70, 765–783.

Fraga, A., I. Goldfajn, and A. Minella (2003): "Inflation Targeting in Emerging Markert Economies," *NBER Macroeconomics Annual*, 18.

Friedman, M. (1960): *A Program for Monetary Stability*, New York: Fordham University Press.

Fudenberg, D. and J. Tirole (1991): *Game Theory*, Cambridge, Mass.: MIT Press.

Galí, J. (2015): *Monetary Policy, Inflation and the Business Cycle: An Introduction to the New Keynesian Framework*, Princeton, N.J., Princeton University Press, second ed.

Garfinkel, M. R. and S. Oh (1993): "Strategic Discipline in Monetary Policy with Private Information: Optimal Targeting Horizons," *American Economic Review*, 83, 99–117.

Gary-Bobo, R. and Y. Spiegel (2006): "Optimal State-Contingent Regulation under Limited Liability," *RAND Journal of Economics*, 37, 431–448.

Giannoni, M. P. and M. Woodford (2017): "Optimal Target Criteria for Stabilization Policy," *Journal of Economic Theory*, 168, 55–106.

Goulder, L. H. and I. W. H. Parry (2008): "Instrument Choice in Environmental Policy," *Review of Environmental Economics and Policy*, 2, 152–174.

Halac, M. and P. Yared (2014): "Fiscal Rules and Discretion under Persistent Shocks," *Econometrica*, 82, 1557–614.

——— (2018): "Fiscal Rules and Discretion in a World Economy," *American Economic Review*, 108, 2305–2334.

——— (2019): "Fiscal Rules and Discretion under Limited Enforcement," Working Paper.

——— (2020): "Inflation Targeting under Political Pressure," Working Paper.

Holmström, B. (1977): "On Incentives and Control in Organizations," Ph.D. Thesis, Stanford University.

——— (1984): "On the Theory of Delegation," in *Bayesian Models in Economic Theory*, ed. by M. Boyer and R. Kihlstrom, North-Holland; New York.

Hüfner, F. (2004): *Foreign Exchange Intervention as a Monetary Policy Instrument - Evidence for Inflation Targeting Countries*, Physica-Verlag Heidelberg.

Innes, R. D. (1990): "Limited Liability and Incentive Contracting with Ex-Ante Action Choices," *Journal of Economic Theory*, 52, 45–67.

International Monetary Fund (2019): *Annual Report on Exchange Arrangements and Exchange Restrictions*, International Monetary Fund.

Kahn, G. (2012): "Estimated Rules for Monetary Policy," *Economic Review, Federal Reserve Bank of Kansas City*, Fourth Quarter.

Kessler, A. S., C. Lülfesmann, and P. W. Schmitz (2005): "Endogenous Punishments in Agency with Verifiable Ex Post Information," *International Economic Review*, 46, 1207–1231.

Kocherlakota, N. (2016): "Rules versus Discretion: A Reconsideration," *Brookings Papers on Economic Activity*, Fall.

Levin, J. (2003): "Relational Incentive Contracts," *American Economic Review*, 93, 835–57.

McCallum, B. (1981): "Price Level Determinacy with an Interest Rate Policy Rule and Rational Expectations," *Journal of Monetary Economics*, 8, 319–329.

McCallum, B. T. and E. Nelson (2005): "Targeting versus Instrument Rules for Monetary Policy," *Federal Reserve Bank of St. Louis Review*, 87, 597–611.

Melumad, N. D. and T. Shibano (1991): "Communication in Settings with No Transfers," *RAND Journal of Economics*, 22, 173–198.

Mirrlees, J. A. (1974): "Notes on Welfare Economics, Information, and Uncertainty," in *Essays on Economic Behavior under Uncertainty*, ed. by Balch, McFadden, and Wu, Amsterdam: North Holland Publishing Co.

Mishkin, F. S. (1999): "International Experiences with Different Monetary Policy Regimes," *Journal of Monetary Economics*, 43, 579–606.

——— (2000): "Inflation Targeting in Emerging-Market Countries," *American Economic Review*, 90, 105–109.

——— (2017): "Improving the Use of Discretion in Monetary Policy," Working Paper.

Nakamura, E., J. Steinsson, P. Sun, and D. Villar (2018): "The Elusive Costs of Inflation: Price Dispersion during the U.S. Great Inflation," *Quarterly Journal of Economics*, 133, 1933–1980.

National Conference of State Legislatures (1999): "State Balanced Budget Requirements," http://www.ncsl.org/research/fiscal-policy/state-balanced-budget-requirements.aspx.

Nikolsko-Rzhevskyy, A., D. H. Papell, and R. Prodan (2019): "The Taylor Principles," *Journal of Macroeconomics*, 62, 103–159.

Padró i Miquel, G. and P. Yared (2012): "The Political Economy of Indirect Control," *Quarterly Journal of Economics*, 127, 947–1015.

Passacantando, F. (1996): "Building an institutional Framework of Monetary Stability: The Case of Italy (1979-1994)," *BNL Quarterly Review, Banca Nazionale del Lavoro*, 49, 83–132.

Poole, W. (1970): "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics*, 84, 197–216.

Prat, A. (2005): "The Wrong Kind of Transparency," *American Economic Review*, 95, 862–877.

Riordan, M. H. and D. E. M. Sappington (1988): "Optimal Contracts with Public ex post Information," *Journal of Economic Theory*, 45, 189–199.

Rogerson, W. (1985): "The First-Order Approach to Principal-Agent Problems," *Econometrica*, 53, 1357–1368.

Rogoff, K. (1985): "The Optimal Degree of Commitment to an Intermediate Monetary Target," *Quarterly Journal of Economics*, 100, 1169–1189.

Romer, C. D. and D. H. Romer (2000): "Federal Reserve Information and the Behavior of Interest Rates," *American Economic Review*, 90, 429–457.

Shapiro, A. H. and D. Wilson (2019): "Taking the Fed at its Word: Direct Estimation of Central Bank Objectives using Text Analytics," Federal Reserve Bank of San Francisco Working Paper 2019-02.

Sleet, C. (2001): "On Credible Monetary Policy and Private Government Information," *Journal of Economic Theory*, 99, 338–376.

Svensson, L. E. (2003): "What is Wrong with Taylor Rules? Using Judgment in Monetary Policy Through Targeting Rules," *Journal of Economic Literature*, 41, 426–477.

——— (2010): "Inflation Targeting," in *Handbook of Monetary Economics*, ed. by B. M. Friedman and M. Woodford, Elsevier, vol. 3B.

Taylor, J. B. (1993): "Discretion versus Policy Rules in Practice," *Carnegie-Rochester Conference Series on Public Policy*, 39, 195–214.

——— (2012): "Monetary Policy Rules Work and Discretion Doesn't: A Tale of Two Eras," *Journal of Money, Credit and Banking*, 44, 1017–1032.

——— (2014): "Inflation Targeting in Emerging Markets: The Global Experience," Speech at the conference on Fourteen Years of Inflation Targeting in South Africa and the Challenge of a Changing Mandate, South African Reserve Bank, Pretoria.

US Congress, Office of Technology Assessment (1995): *Environmental Policy Tools: A User's Guide*, Washington, DC: U.S. Government Printing Office (OTA-ENV-634).

Waki, Y., R. Dennis, and I. Fujiwara (2018): "The Optimal Degree of Monetary-Discretion in a New Keynesian Model with Private Information," *Theoretical Economics*, 13, 1319–1368.

Walsh, C. (2015): "Goals and Rules in Central Bank Design," *International Journal of Central Banking*, 11, 295–352.

Walsh, C. E. (1995): "Optimal Contracts for Central Bankers," *American Economic Review*, 85, 150–167.

——— (2002): "When Should Central Bankers Be Fired?" *Economics of Governance*, 3, 1–21.

Weitzman, M. L. (1974): "Prices vs. Quantities," *Review of Economic Studies*, 41, 477–491.

Woodford, M. (2003): *Interest and Prices: Foundations of a Theory of Monetary Policy*, Princeton, N.J., Princeton University Press.

——— (2007): "The Case for Forecast Targeting as a Monetary Policy Strategy," *Journal of Economic Perspectives*, 21, 3–24.

Yellen, J. L. (2012): "The Economic Outlook and Monetary Policy," Speech at the Money Marketeers of New York University, New York.

——— (2015): "Normalizing Monetary Policy: Prospects and Perspectives," Speech at the The New Normal Monetary Policy Conference, `https://www.federalreserve.gov/newsevents/speech/yellen20150327a.htm`.

——— (2017): "The Economic Outlook and the Conduct of Monetary Policy," Speech at At the SIEPR, Stanford University, `https://www.federalreserve.gov/newsevents/speech/yellen20170119a.htm`.

A Proofs for Section 3 and Section 4

This Appendix provides proofs for the results in Section 3 and Section 4. See the Online Appendix for proofs of the results in Section 5.

A.1 Proof of Proposition 1

We proceed in three steps.

Step 1. We solve a relaxed version of the program in (12) which ignores (8) for $j = H$ and (9) for $j = L, H$. Step 2 verifies that the solution to this relaxed problem satisfies these constraints.

Step 1a. We show that the solution to the relaxed problem satisfies (8) for $j = L$ as an equality. If this were not the case, then the solution would set $\mu^j = s^j$ and $P^j = 0$ for $j = L, H$. However, (8) for $j = L$ would then become

$$\int_{-\infty}^{\infty} \left[\alpha s^L + U(s^L, \theta, 0)\right] \phi(\theta|s^L, \sigma^2) d\theta \geq \int_{-\infty}^{\infty} \left[\alpha s^H + U(s^H, \theta, 0)\right] \phi(\theta|s^L, \sigma^2) d\theta,$$

which after some algebra yields

$$\left(s^H - s^L\right)\left(s^H - s^L - 2\alpha\right) \geq 0.$$

This inequality contradicts Assumption 1. Thus, (8) for $j = L$ must bind.

Step 1b. We show that the solution to the relaxed problem satisfies $\mu^H \geq \mu^L$. Suppose by contradiction that $\mu^H < \mu^L$. Consider two perturbations, one assigning μ^L and no punishment to both types, and another assigning μ^H no punishment to both types. Since these perturbations are feasible and incentive compatible, the contradiction assumption requires that neither of them strictly increase social welfare, which requires:

$$\sum_{j=L,H} \frac{1}{2} \int_{-\infty}^{\infty} U(\mu^j, \theta, \mathbb{E}\mu)\phi(\theta|s^j, \sigma^2)d\theta \geq \sum_{j=L,H} \frac{1}{2} \int_{-\infty}^{\infty} U(\mu^L, \theta, \mu^L)\phi(\theta|s^j, \sigma^2)d\theta, \text{ and}$$
$$\sum_{j=L,H} \frac{1}{2} \int_{-\infty}^{\infty} U(\mu^j, \theta, \mathbb{E}\mu)\phi(\theta|s^j, \sigma^2)d\theta \geq \sum_{j=L,H} \frac{1}{2} \int_{-\infty}^{\infty} U(\mu^H, \theta, \mu^H)\phi(\theta|s^j, \sigma^2)d\theta.$$

Let $\omega \equiv \beta\left(\frac{\mu^L - \mu^H}{2}\right)$. Given $\mathbb{E}\mu = \frac{\mu^L + \mu^H}{2}$, these two inequalities can be rewritten as

$$\sum_{j=L,H} \int_{-\infty}^{\infty} \left\{U(\mu^j, \theta, \mathbb{E}\mu) - U(\mu^L, \theta, \mathbb{E}\mu) + \frac{\gamma}{2}\left[\omega^2 - 2(\mu^L - \theta - \beta\mathbb{E}\mu)\omega\right]\right\} \phi(\theta|s^j, \sigma^2)d\theta \geq 0, \text{ and}$$
$$\sum_{j=L,H} \int_{-\infty}^{\infty} \left\{U(\mu^j, \theta, \mathbb{E}\mu) - U(\mu^H, \theta, \mathbb{E}\mu) + \frac{\gamma}{2}\left[\omega^2 + 2(\mu^H - \theta - \beta\mathbb{E}\mu)\omega\right]\right\} \phi(\theta|s^j, \sigma^2)d\theta \geq 0.$$

Adding these two inequalities and rearranging terms yields

$$\int_{-\infty}^{\infty} \left[U(\mu^H, \theta, \mathbb{E}\mu) - U(\mu^L, \theta, \mathbb{E}\mu)\right]\left[\phi(\theta|s^H, \sigma^2) - \phi(\theta|s^L, \sigma^2)\right] d\theta + 2\gamma\omega[\omega - (\mu^L - \mu^H)] \geq 0.$$

However, since $s^H > s^L$, both terms on the left-hand side are strictly negative if $\mu^H < \mu^L$, yielding a contradiction. Hence, $\mu^H \geq \mu^L$.

Step 1c. We show that the solution to the relaxed problem satisfies $P^L = P^H = 0$. Note first that if $P^L > 0$, then a reduction in P^L is feasible, relaxes constraint (8) for $j = L$, and strictly increases the objective. Hence, $P^L = 0$, and therefore (8) for $j = L$ (which binds by Step 1a) can be rewritten as:

$$\left(\alpha + s^L + \gamma\beta\mathbb{E}\mu\right)\mu^L - \frac{(\mu^L)^2}{2} = \left(\alpha + s^L + \gamma\beta\mathbb{E}\mu\right)\mu^H - \frac{(\mu^H)^2}{2} - P^H. \tag{28}$$

This equation implies that, up to an additive constant independent of the allocation, the high type's welfare satisfies

$$\left(\alpha+s^{H}+\gamma\beta\mathbb{E}\mu\right)\mu^{H}-\frac{(\mu^{H})^{2}}{2}-P^{H}=\left(\alpha+s^{L}+\gamma\beta\mathbb{E}\mu\right)\mu^{L}-\frac{(\mu^{L})^{2}}{2}+\left(s^{H}-s^{L}\right)\mu^{H}. \tag{29}$$

Now suppose by contradiction that $P^H > 0$. It follows from (28) and Step 1b that $\mu^H > \mu^L$. Note that up to an additive constant independent of the allocation, social welfare in (12) is equal to

$$\frac{1}{2}\left[\left(s^{L}+\gamma\beta\mathbb{E}\mu\right)\mu^{L}-\frac{(\mu^{L})^{2}}{2}\right]+\frac{1}{2}\left[\left(s^{H}+\gamma\beta\mathbb{E}\mu\right)\mu^{H}-\frac{(\mu^{H})^{2}}{2}-P^{H}\right]-\frac{\gamma}{2}(\beta\mathbb{E}\mu)^{2}.$$

Substituting with (29) and $\mathbb{E}\mu = \frac{\mu^L+\mu^H}{2}$, multiplying by 2, and rearranging terms yields

$$\left(\frac{1}{2}\alpha+s^{L}\right)\mu^{L}-\frac{(\mu^{L})^{2}}{2}+\gamma\beta\left[\mu^{L}\left(\frac{\mu^{L}+\mu^{H}}{2}\right)-\frac{\beta}{2}\left(\frac{\mu^{L}+\mu^{H}}{2}\right)^{2}\right]+\frac{1}{2}\left(s^{H}-s^{L}-\alpha\right)\mu^{H}. \tag{30}$$

The first-order conditions for μ^L and μ^H respectively are:

$$\begin{aligned}\left(\alpha+2s^{L}\right)-(2-\gamma\beta)\mu^{L}+\gamma\beta\left(1-\frac{\beta}{2}\right)\left(\mu^{L}+\mu^{H}\right) &= 0, &(31)\\ \gamma\beta\mu^{L}-\gamma\beta^{2}\frac{\left(\mu^{L}+\mu^{H}\right)}{2}+\left(s^{H}-s^{L}-\alpha\right) &= 0. &(32)\end{aligned}$$

We establish that these two conditions cannot hold. By Assumption 1, the last term in (32) is negative. Therefore, (32) requires

$$\gamma\beta\mu^{L}-\gamma\beta^{2}\frac{\left(\mu^{L}+\mu^{H}\right)}{2}\geq 0,$$

which in turn requires $\mu^L > 0$, since $\mu^H > \mu^L$ and $\beta \in (0, 1)$. However, combining (31) and (32) implies

$$\mu^{L}=\frac{(1-\beta)}{\beta(1-\gamma)}(s^{H}-s^{L}-\alpha), \tag{33}$$

which implies $\mu^L \leq 0$ by Assumption 1 and thus yields a contradiction. It follows that $P^H > 0$ cannot hold, so $P^H = 0$ in the solution.

Step 1d. We show that the solution to the relaxed problem satisfies $\mu^L = \mu^H = 0$. By Step 1b, if $\mu^L \neq \mu^H$, then $\mu^H > \mu^L$. However, the arguments in Step 1c imply that $\mu^H > \mu^L$ cannot hold. Therefore, $\mu^H = \mu^L = \mu$ for some μ. Social welfare in (30) is thus equal to

$$-\left[\frac{1}{2} - \gamma\beta\left(1 - \frac{\beta}{2}\right)\right]\mu^2,$$

which is strictly decreasing in μ^2 given $\gamma \in [0,1]$ and $\beta \in (0,1)$. Therefore, social welfare is maximized at $\mu^L = \mu^H = 0$.

Step 2. We verify that the solution to the relaxed problem in Step 1 satisfies the constraints of the original problem. Since $\mu^L = \mu^H$ and $P^L = P^H$, constraint (8) for $j = H$ is satisfied. As for the constraints in (9), given $\mu^L = \mu^H = 0$, these require, for $j = L, H$,

$$\overline{P} \geq \frac{(\alpha + s^j)^2}{2}.$$

By Assumption 1, this inequality holds for $j = L, H$ if

$$\overline{P} \geq \frac{9}{8}\alpha^2,$$

which is satisfied by Assumption 2 and the fact that $2\phi(1|0,1) < 0.5$.

Step 3. We verify that a maximally-enforced instrument threshold $\mu^* = 0$ implements the solution. Given (6) and (7), conditional on choosing a policy $\mu \leq \mu^*$ and receiving no punishment, the central bank's optimal choice is $\mu = \mu^*$ regardless of type. Moreover, conditional on choosing a policy $\mu > \mu^*$ and receiving maximal punishment $\overline{P}$, the central bank's optimal choice is $s^j + \alpha + \gamma\beta\mathbb{E}\mu$ for each $j = L, H$. The enforcement constraints in (9) guarantee that the central bank has no incentive to deviate to $\mu > \mu^*$.

A.2 Proof of Proposition 2

We proceed in two steps.

Step 1. We follow a first-order approach by solving a relaxed version of the program in (16) that replaces (13) with the central bank's first-order condition (18). Step 2 verifies the validity of this approach.

As noted in the text, the solution to (18) is $\mu^j = s^j + \delta$ for $j = L, H$ and some $\delta \gtreqless 0$. Hence, the relaxed problem, up to an additive constant independent of the allocation, can be written as:

$$\max_{\delta,\{P(\pi)\}_{\pi\in\mathbb{R}}} \left\{ -[1-\gamma\beta(2-\beta)]\frac{\delta^2}{2} - \int_{-\infty}^{\infty} P(\pi)\,\phi(\delta-\pi|0,\sigma^2)d\pi \right\} \tag{34}$$

subject to

$$\alpha - (1-\gamma\beta)\delta - \int_{-\infty}^{\infty} P(\pi)\phi'\left(\delta-\pi|0,\sigma^2\right)d\pi = 0, \tag{35}$$

$$P(\pi) \in \left[0, \overline{P}\right] \text{ for all } \pi \in \mathbb{R}. \tag{36}$$

Step 1a. Denote by λ the Lagrange multiplier on (35). We show that $\lambda < 0$. To do this, we consider a doubly-relaxed problem in which constraint (35) is replaced with an inequality constraint (cf. Rogerson, 1985):

$$\alpha - (1-\gamma\beta)\delta - \int_{-\infty}^{\infty} P(\pi)\phi'\left(\delta-\pi|0,\sigma^2\right)d\pi \le 0. \tag{37}$$

Since this is an inequality constraint, the multiplier satisfies $\lambda \le 0$. We show that (37) holds as an equality in the solution to the doubly-relaxed problem, and thus this problem is equivalent to (34)-(36) with $\lambda < 0$. Suppose by contradiction that (37) holds as a strict inequality. Then to maximize (34) the principal chooses $\delta = 0$ and $P(\pi) = 0$ for all π. However, substituting back into the left-hand side of (37) yields $\alpha \le 0$, which is a contradiction since $\alpha > 0$. Therefore, (37) holds as an equality in the doubly-relaxed problem and $\lambda < 0$.

Step 1b. We show that the solution to (34)-(36) satisfies $P(\pi) = 0$ if $\pi \le \pi^*$ and $P(\pi) = \overline{P}$ if $\pi > \pi^*$, for some $\pi^* \in (-\infty, \infty)$. Denote by $\overline{\psi}(\pi)$ the Lagrange multiplier on $P(\pi) \le \overline{P}$ and, analogously, denote by $\underline{\psi}(\pi)$ the Lagrange multiplier on $P(\pi) \ge 0$. The first-order condition with respect to $P(\pi)$ is

$$-\phi(\delta-\pi|0,\sigma^2) - \lambda\phi'(\delta-\pi|0,\sigma^2) + \underline{\psi}(\pi) - \overline{\psi}(\pi) = 0. \tag{38}$$

Suppose that $P(\pi)$ is strictly interior with $\underline{\psi}(\pi) = \overline{\psi}(\pi) = 0$. Then (38) yields

$$-\frac{1}{\lambda} = \frac{\phi'(\delta - \pi|0, \sigma^2)}{\phi(\delta - \pi|0, \sigma^2)} = \frac{\pi - \delta}{\sigma^2}. \tag{39}$$

Since the right-hand side of (39) is strictly increasing in π whereas the left-hand side is constant, it follows that (39) holds for only one value of $\pi \in (-\infty, \infty)$, which we label π^*. By (38) and (39), the solution has $P(\pi) = 0$ if $\pi \leq \pi^*$ and $P(\pi) = \overline{P}$ if $\pi > \pi^*$.

Step 1c. We show that $\pi^* > \delta$ and $\delta \in \left(0, \frac{\alpha}{1-\gamma\beta}\right)$. Moreover, $\pi^* = \delta + \frac{\sigma^2(1-\gamma\beta)}{\delta[1-\gamma\beta(2-\beta)]}$.

To show $\pi^* > \delta$, recall from Step 1a that $\lambda < 0$; hence, (39) yields $\pi^* > \delta$. To show $\delta < \frac{\alpha}{1-\gamma\beta}$, note that by Step 1b, (35) can be rewritten as

$$\alpha - (1-\gamma\beta)\delta - \phi\left(\delta - \pi^*|0, \sigma^2\right)\overline{P} = 0. \tag{40}$$

Since $\phi\left(\delta - \pi^*|0, \sigma^2\right)\overline{P} > 0$, (40) requires $\alpha > (1-\gamma\beta)\delta$.

We next show $\delta > 0$. By Step 1b, the Lagrangian to solve for the socially optimal levels of δ and π^* can be written as

$$-[1-\gamma\beta(2-\beta)]\frac{\delta^2}{2} - \Phi(\delta - \pi^*|0, \sigma^2)\overline{P} + \lambda\left[\alpha - (1-\gamma\beta)\delta - \phi\left(\delta - \pi^*|0, \sigma^2\right)\overline{P}\right]. \tag{41}$$

The first-order condition with respect to δ is

$$-[1-\gamma\beta(2-\beta)]\delta - \phi(\delta - \pi^*|0, \sigma^2)\overline{P} - \lambda\left[1 - \gamma\beta + \phi'(\delta - \pi^*|0, \sigma^2)\overline{P}\right] = 0, \tag{42}$$

and the first-order condition with respect to π^* is

$$\phi(\delta - \pi^*|0, \sigma^2)\overline{P} + \lambda\phi'(\delta - \pi^*|0, \sigma^2)\overline{P} = 0. \tag{43}$$

Substituting (43) into (42) yields

$$-\lambda = \frac{[1-\gamma\beta(2-\beta)]}{1-\gamma\beta}\delta. \tag{44}$$

Since $\lambda < 0$ by Step 1a, $\gamma \in [0, 1]$, and $\beta \in (0, 1)$, (44) implies $\delta > 0$.

Finally, we show that $\pi^* = \delta + \frac{\sigma^2(1-\gamma\beta)}{\delta[1-\gamma\beta(2-\beta)]}$. This follows from combining (39) and

(44), which yields $\delta - \pi^* = -\frac{\sigma^2}{\delta\chi}$ for

$$\chi \equiv \frac{1-\gamma\beta(2-\beta)}{1-\gamma\beta}. \tag{45}$$

Step 2. We verify the validity of the first-order approach: we establish that the choice of δ in the relaxed problem satisfies (13) and therefore corresponds to the central bank's global optimum.

Step 2a. We begin by showing that the central bank has no incentive to choose some $\delta' \neq \delta$, $\delta' \leq \pi^*$. The central bank's first-order condition (40) which takes $\mathbb{E}\mu = \delta^e$ as given can be rewritten as

$$\alpha - \delta + \gamma\beta\delta^e - \phi\left(\delta - \pi^*|0,\sigma^2\right)\overline{P} = 0.$$

Differentiating the first-order condition with respect to δ yields

$$-1 - \phi'(\delta - \pi^*|0,\sigma^2)\overline{P}. \tag{46}$$

Note that (46) is strictly negative for all $\delta \leq \pi^*$, and thus the central bank's welfare is strictly concave over this range. Since by Step 1c the solution to the relaxed problem sets $\delta < \pi^*$, we conclude that this δ is a maximum and dominates any other $\delta' \leq \pi^*$.

Step 2b. We next show that the central bank has no incentive to choose some $\delta' \neq \delta$, $\delta' > \pi^*$. To prove this, we first establish that in the solution to the relaxed problem, given π^*, δ satisfies $\delta - \pi^* \leq -\sigma$. Suppose by contradiction that $\delta - \pi^* > -\sigma$. Recall from Step 1c that $\delta - \pi^* = -\frac{\sigma^2}{\delta\chi}$ for χ defined in (45). Hence, the contradiction assumption implies $\delta\chi > \sigma$. Substituting $\delta - \pi^* = -\frac{\sigma^2}{\delta\chi}$ into (40) yields

$$\frac{\alpha}{1-\gamma\beta} - \delta - \phi\left(-\frac{\sigma^2}{\delta\chi}|0,\sigma^2\right)\frac{\overline{P}}{1-\gamma\beta} = 0. \tag{47}$$

Since the left-hand side of (47) is decreasing in δ and (by the contradiction assumption) $\delta\chi > \sigma$, (47) requires

$$\frac{\alpha}{1-\gamma\beta} - \frac{\sigma}{\chi} - \phi\left(-\sigma|0,\sigma^2\right)\frac{\overline{P}}{1-\gamma\beta} > 0.$$

Multiply both sides of this equation by $\frac{\sigma}{\chi} > 0$ to obtain

$$\frac{\sigma}{\chi}\left(\frac{\alpha}{1-\gamma\beta} - \frac{\sigma}{\chi}\right) - \sigma\phi\left(-\sigma|0,\sigma^2\right)\frac{\overline{P}}{1-\gamma\beta(2-\beta)} > 0. \tag{48}$$

Note that $\chi > 1-\gamma\beta$. Moreover, since $0 < \sigma < \sqrt{Var(\theta)}$ and, by Assumption 1, $\sqrt{Var(\theta)} \leq \frac{\alpha}{2}$, we have $\sigma < \frac{\alpha}{2}$. Thus, we obtain

$$\frac{\sigma}{\chi}\left(\frac{\alpha}{1-\gamma\beta} - \frac{\sigma}{\chi}\right) < \frac{1}{2}\frac{\alpha^2}{(1-\gamma\beta)^2},$$

and therefore (48) yields

$$\frac{1}{2\sigma\phi\left(-\sigma|0,\sigma^2\right)}\frac{\alpha^2[1-\gamma\beta(2-\beta)]}{(1-\gamma\beta)^2} > \overline{P}.$$

However, this inequality violates Assumption 2 since $\sigma\phi\left(-\sigma|0,\sigma^2\right) = \phi\left(1|0,1\right)$. Thus, given π^*, δ satisfies $\delta - \pi^* \leq -\sigma$.

We can now establish that the central bank has no incentive to deviate to $\delta' \neq \delta$, $\delta' > \pi^*$. Consider some $\delta' > \pi^*$ that is a local maximum for the central bank. The difference in welfare for the central bank from choosing the value of δ given by the solution to the relaxed problem versus δ' is equal to

$$(\alpha + \gamma\beta\delta^e)(\delta - \delta') - \left(\frac{\delta^2}{2} - \frac{(\delta')^2}{2}\right) - \left[\Phi\left(\delta - \pi^*|0,\sigma^2\right) - \Phi\left(\delta' - \pi^*|0,\sigma^2\right)\right]\overline{P}. \tag{49}$$

Note that by the arguments in Step 1c and δ and δ' satisfying the central bank's first-order condition, it follows that both δ and δ' are between 0 and $\alpha + \gamma\beta\delta^e$. Thus, (49) is bounded from below by

$$-\frac{1}{2}\left(\alpha + \gamma\beta\delta^e\right)^2 - \left[\Phi\left(\delta - \pi^*|0,\sigma^2\right) - \Phi\left(\delta' - \pi^*|0,\sigma^2\right)\right]\overline{P}. \tag{50}$$

Moreover, from Step 1c, δ^e is positive and below $\alpha/(1-\gamma\beta)$. Therefore, (50) is bounded from below by

$$-\frac{1}{2}\left(\frac{\alpha}{1-\gamma\beta}\right)^2 - \left[\Phi\left(\delta - \pi^*|0,\sigma^2\right) - \Phi\left(\delta' - \pi^*|0,\sigma^2\right)\right]\overline{P}. \tag{51}$$

Since (40) is satisfied for both δ and δ' and $\delta' > \delta$, we must have $\phi\left(\delta - \pi^*|0,\sigma^2\right) >$

$\phi(\delta' - \pi^*|0, \sigma^2)$. Moreover, by the symmetry of the normal distribution, $\phi(\delta - \pi^*|0, \sigma^2) = \phi(-(\delta - \pi^*)|0, \sigma^2)$ and thus $\Phi(-(\delta - \pi^*)|0, \sigma^2) < \Phi(\delta' - \pi^*|0, \sigma^2)$. Therefore, (51) is bounded from below by

$$-\frac{1}{2}\left(\frac{\alpha}{1-\gamma\beta}\right)^2 + \left[\Phi\left(-(\delta - \pi^*)|0, \sigma^2\right) - \Phi\left(\delta - \pi^*|0, \sigma^2\right)\right]\overline{P}. \tag{52}$$

Since, as shown above, $\delta - \pi^* \leq -\sigma$, we obtain that (52) is itself bounded from below by

$$\begin{aligned}
&-\frac{1}{2}\left(\frac{\alpha}{1-\gamma\beta}\right)^2 + \left[\Phi\left(\sigma|0, \sigma^2\right) - \Phi\left(-\sigma|0, \sigma^2\right)\right]\overline{P} \\
&= -\frac{1}{2}\left(\frac{\alpha}{1-\gamma\beta}\right)^2 + \left[\Phi(1|0,1) - \Phi(-1|0,1)\right]\overline{P} \\
&> 0,
\end{aligned}$$

where the last inequality follows from Assumption 2 and the fact that $\phi(1|0,1) < \Phi(1|0,1) - \Phi(-1|0,1)$. Therefore, the central bank strictly prefers δ over δ'.

A.3 Proof of Proposition 3

We begin by proving the following lemma:

Lemma 1. *Consider changing σ while keeping $Var(\theta)$ unchanged. Social welfare is independent of σ under the optimal instrument-based rule and it is strictly decreasing in σ under the optimal target-based rule.*

Proof. By Proposition 1, an optimal instrument-based rule sets $\mu^j = 0$ and $P^j = 0$ for $j = L, H$. Since $Var(\theta) = \mathbb{E}(\theta^2) - (\mathbb{E}(\theta))^2 = \mathbb{E}(\theta^2)$ (by $\mathbb{E}(\theta) = 0$), social welfare under this rule is equal to $-\frac{Var(\theta)}{2}$, which is independent of σ.

To evaluate social welfare under an optimal target-based rule, consider the Lagrangian taking into account the conditional variance term (which is exogenous and thus excluded from (41)):

$$-\frac{\sigma^2}{2} - [1 - \gamma\beta(2-\beta)]\frac{\delta^2}{2} - \Phi(\delta - \pi^*|0, \sigma^2)\overline{P} + \lambda\left[\alpha - (1-\gamma\beta)\delta - \phi(\delta - \pi^*|0, \sigma^2)\overline{P}\right].$$

The derivative with respect to σ is:

$$-\sigma + \overline{P}\left[\int_{\delta-\pi^*}^{\infty}\left(-\frac{\sigma^2 - z^2}{\sigma^3}\right)\phi(z|0,\sigma^2)dz + \lambda\frac{\sigma^2 - (\delta-\pi^*)^2}{\sigma^3}\phi(\delta-\pi^*|0,\sigma^2)\right].$$

The first term is strictly negative. Using (39) and (44) to substitute in for λ and π^*, the sign of the second term is the same as the sign of

$$-\int_{-\frac{\sigma^2}{\delta\chi}}^{\infty}\left(\sigma^2 - z^2\right)\phi(z|0,\sigma^2)dz - \delta\chi\left[\sigma^2 - \left(\frac{\sigma^2}{\delta\chi}\right)^2\right]\phi\left(-\frac{\sigma^2}{\delta\chi}|0,\sigma^2\right), \tag{53}$$

where χ is defined in (45). We next show that this expression is strictly negative, which proves the claim. Consider the derivative of (53) with respect to δ:

$$\left[\left(\sigma^2 - \left(\frac{\sigma^2}{\delta\chi}\right)^2\right)\frac{\sigma^2}{(\delta\chi)^2} - \left(\sigma^2 - \left(\frac{\sigma^2}{\delta\chi}\right)^2\right) - 2\left(\frac{\sigma^2}{\delta\chi}\right)^2 - \left(\sigma^2 - \left(\frac{\sigma^2}{\delta\chi}\right)^2\right)\frac{\sigma^2}{(\delta\chi)^2}\right]\phi\left(-\frac{\sigma^2}{\delta\chi}|0,\sigma^2\right).$$

This derivative takes the same sign as

$$-\left(\sigma^2 - \left(\frac{\sigma^2}{\delta\chi}\right)^2\right) - 2\left(\frac{\sigma^2}{\delta\chi}\right)^2,$$

which is strictly negative. Hence, since $\delta > 0$, it suffices to show that the sign of (53) is weakly negative for $\delta \to 0$. By the definition of variance, the first term in (53) goes to zero as $\delta \to 0$. The second term in (53) can be rewritten as:

$$-\sigma^2\delta\chi\phi\left(-\frac{\sigma^2}{\delta\chi}|0,\sigma^2\right) + \frac{\sigma^4}{\delta\chi}\phi\left(-\frac{\sigma^2}{\delta\chi}|0,\sigma^2\right). \tag{54}$$

As $\delta \to 0$, the first term in (54) goes to zero. Moreover, applying L'Hopital's Rule on

$$\frac{1/(\delta\chi)}{\phi\left(-\frac{\sigma^2}{\delta\chi}|0,\sigma^2\right)^{-1}}$$

shows that the second term also goes to zero. □

We now proceed with the proof of Proposition 3. By Lemma 1, social welfare under the optimal instrument-based rule is invariant to σ, whereas social welfare under the optimal target-based rule is decreasing in σ. To prove the first part of the proposition, it thus suffices to show that a target-based rule is socially optimal at one extreme, for $\sigma \to 0$, and an instrument-based rule is socially optimal at the other extreme, for $\sigma \to \sqrt{Var(\theta)}$. This is what we prove next.

Consider first the case of $\sigma \to 0$. By the arguments in Step 1c and Step 2b of the proof of Proposition 2, $0 < \delta\chi \leq \sigma$. Hence, $\delta \to 0$ as $\sigma \to 0$. Moreover, as $\sigma \to 0$, $\phi\left(z|0,\sigma^2\right)$ corresponds to a Dirac's delta function, with cumulative distribution function $\Phi(z|0,\sigma^2) = 0$ if $z < 0$ and $\Phi(z|0,\sigma^2) = 1$ if $z \geq 0$. Therefore, since $\delta - \pi^* < 0$ in the optimal target-based rule, the limit of social welfare under this rule, as $\sigma \to 0$, is given by

$$\lim_{\sigma \to 0} \left\{ -[1 - \gamma\beta(2-\beta)]\frac{\delta^2}{2} - \Phi\left(\delta - \pi^*|0,\sigma^2\right)\overline{P} \right\} = 0.$$

Since social welfare under the optimal instrument-based rule is $-\frac{Var(\theta)}{2}$, it follows that the optimal target-based rule dominates the optimal instrument-based rule.

Consider next the case of $\sigma \to \sqrt{Var\left(\theta\right)}$ and thus $\Delta \to 0$. Since δ in the optimal target-based rule satisfies equation (47), the solution in this case admits $\delta > 0$. Social welfare under the optimal target-based rule is then equal to

$$-\frac{Var(\theta)}{2} - [1 - \gamma\beta(2-\beta)]\frac{\delta^2}{2} - \Phi\left(\delta - \pi^*|0,\sigma^2\right)\overline{P}.$$

Since this is strictly lower than $-\frac{Var(\theta)}{2}$, it follows that the optimal instrument-based rule dominates the optimal target-based rule.

Finally, to prove the second part of the proposition, note that social welfare under the optimal instrument-based rule is independent of the inflation bias α and the punishment $\overline{P}$. Thus, it suffices to show that social welfare under the optimal target-based rule is decreasing in α and increasing in $\overline{P}$. The former follows from the fact that the derivative of the Lagrangian in (41) with respect to α is equal to λ, which is strictly negative by Step 1a in the proof of Proposition 2. To evaluate how welfare changes with $\overline{P}$, consider the representation of the program in (16). An increase in $\overline{P}$ relaxes constraint (14). Since this constraint is binding in the solution (by Step 1b of the proof of Proposition 2), it follows that an increase in $\overline{P}$ strictly increases social welfare under the optimal target-based rule.

A.4 Proof of Proposition 4

We proceed in three steps.

Step 1. We solve a relaxed version of (19)-(22) which ignores (20) for $j = H$ and (21) for $j = L, H$. Step 2 verifies that the solution to this relaxed problem satisfies these constraints.

Step 1a. We show that the solution satisfies (20) for $j = L$ as an equality. The proof of this claim is analogous to that in Step 1a of the proof of Proposition 1 and thus omitted.

Step 1b. We show that the solution satisfies $\mu^H \geq \mu^L$. The proof of this claim is analogous to that in Step 1b of the proof of Proposition 1 and thus omitted.

Step 1c. We show that the solution satisfies $P^L(\theta) = 0$ for all θ. If $P^L(\theta) > 0$ for some θ, then a decrease in $P^L(\theta)$ is feasible, relaxes constraint (20) for $j = L$, and strictly increases the objective. The claim follows.

Step 1d. We show that the solution satisfies $P^H(\theta) = \overline{P}$ if $\theta < \theta^*$ and $P^H(\theta) = 0$ if $\theta \geq \theta^*$, for some $\theta^* \in (-\infty, \infty)$. Let $\frac{1}{2}\lambda$ be the Lagrange multiplier on (20) and denote by $\overline{\psi}(\theta)$ and $\underline{\psi}(\theta)$ the Lagrange multipliers on $P^H(\theta) \leq \overline{P}$ and $P^H(\theta) \geq 0$ respectively. The first-order condition with respect to $P^H(\theta)$ yields

$$-\frac{1}{2}\phi(\theta|s^H, \sigma^2) + \frac{1}{2}\lambda\phi(\theta|s^L, \sigma^2) + \underline{\psi}(\theta) - \overline{\psi}(\theta) = 0. \tag{55}$$

Suppose that $P^H(\theta)$ is strictly interior with $\underline{\psi}(\theta) = \overline{\psi}(\theta) = 0$. Then (55) implies

$$\lambda = \frac{\phi\left(\theta - s^H|0, \sigma^2\right)}{\phi\left(\theta - s^L|0, \sigma^2\right)}. \tag{56}$$

Since the right-hand side of (56) is strictly increasing in θ whereas the left-hand side is constant, it follows that (56) holds only for one value of $\theta \in (-\infty, \infty)$, which we label θ^*. By (55) and (56), the solution has $P^H(\theta) = \overline{P}$ if $\theta < \theta^*$ and $P^H(\theta) = 0$ if $\theta \geq \theta^*$.

Step 1e. We show that the solution admits $\lambda < 1$ and, as a consequence, $\theta^* < 0$. Note that, given Assumption 1, $\lambda < 1$ is implied if the following inequality holds:

$$-\lambda\alpha + (1 + \lambda)\Delta \geq \lambda\alpha - (1 + \lambda)\Delta. \tag{57}$$

We therefore proceed by proving that (57) holds. Suppose by contradiction that (57) does not hold. By Steps 1a, 1c, and 1d, constraint (20) for $j = L$ can be written as

$$\left(\alpha + s^L + \gamma\beta\mathbb{E}\mu\right)\mu^L - \frac{(\mu^L)^2}{2} = \left(\alpha + s^L + \gamma\beta\mathbb{E}\mu\right)\mu^H - \frac{(\mu^H)^2}{2} - \Phi\left(\theta^* - s^L|0,\sigma^2\right)\overline{P}.$$

Hence, the Lagrangian for the relaxed problem can be represented as

$$\frac{1}{2}\left[s^L\mu^L - \frac{(\mu^L)^2}{2}\right] + \frac{1}{2}\left[s^H\mu^H - \frac{(\mu^H)^2}{2}\right] + \gamma\beta\left(1 - \frac{\beta}{2}\right)(\mathbb{E}\mu)^2 - \frac{1}{2}\Phi\left(\theta^* - s^H|0,\sigma^2\right)\overline{P}$$
$$+\frac{\lambda}{2}\left[\left(\alpha + s^L + \gamma\beta\mathbb{E}\mu\right)\left(\mu^L - \mu^H\right) - \frac{\left(\mu^L\right)^2 - \left(\mu^H\right)^2}{2} + \Phi\left(\theta^* - s^L|0,\sigma^2\right)\overline{P}\right].$$

The first-order conditions with respect to μ^L and μ^H yield

$$-\mu^L[1+\lambda(1-\gamma\beta)] + \gamma\beta\,(2-\beta)\,\mathbb{E}\mu \;=\; -\lambda\alpha + (1+\lambda)\Delta, \text{ and} \tag{58}$$
$$-\mu^H[1-\lambda(1-\gamma\beta)] + \gamma\beta\,(2-\beta)\,\mathbb{E}\mu \;=\; \lambda\alpha - (1+\lambda)\Delta. \tag{59}$$

The second order condition with respect to μ^H yields

$$-\left[1-\lambda(1-\gamma\beta)\right] + \frac{\gamma\beta(2-\beta)}{2} < 0, \tag{60}$$

which implies

$$0 < 1-\lambda(1-\gamma\beta) < 1+\lambda(1-\gamma\beta). \tag{61}$$

Note that by the previous steps, $\mu^L < \mathbb{E}\mu < \mu^H$. Using this and (61), the first-order conditions (58) and (59) imply

$$\mathbb{E}\mu\left\{\gamma\beta\,(2-\beta) - [1+\lambda(1-\gamma\beta)]\right\} \;<\; -\lambda\alpha + (1+\lambda)\Delta, \text{ and} \tag{62}$$
$$\mathbb{E}\mu\left\{\gamma\beta\,(2-\beta) - [1-\lambda(1-\gamma\beta)]\right\} \;>\; \lambda\alpha - (1+\lambda)\Delta. \tag{63}$$

Recall that by the contradiction assumption, (57) is violated. Hence, given (61), the inequalities (62) and (63) can only hold if $\mathbb{E}\mu > 0$. However, if (57) is violated, then (58) and (59) can be combined to yield

$$-\mu^L[1+\lambda(1-\gamma\beta)] < -\mu^H[1-\lambda(1-\gamma\beta)]. \tag{64}$$

Given $\mu^H > \mu^L$ and (61), this requires $\mu^L < 0$ and $\mu^H < 0$, implying $\mathbb{E}\mu < 0$ and thus yielding a contradiction. Therefore, we obtain that (57) holds and thus $\lambda < 1$.

We end this step by observing that since $\lambda < 1$ and $s^H = -s^L = \Delta$, (56) implies $\theta^* < 0$.

Step 1f. We show that the solution satisfies $\mu^H \leq \Delta + \gamma\beta\mathbb{E}\mu$, which implies $\mu^L < \mu^f(s^L, \mathbb{E}\mu) = \alpha + s^L + \gamma\beta\mathbb{E}\mu$ (by $\mu^L < \mu^H$ and Assumption 1) and $\mu^H < \mu^f(s^H, \mathbb{E}\mu) = \alpha + s^H + \gamma\beta\mathbb{E}\mu$. Suppose by contradiction that $\mu^H > \Delta + \gamma\beta\mathbb{E}\mu$. Adding (58) and (59) yields

$$-\left[1 - \gamma\beta\left(2-\beta\right)\right]2\mathbb{E}\mu + \lambda\left(1-\gamma\beta\right)\left(\mu^H - \mu^L\right) = 0. \tag{65}$$

Subtracting (59) from (58) yields:

$$\mu^H - \mu^L = \lambda(1-\gamma\beta)2\mathbb{E}\mu + 2[(1+\lambda)\Delta - \lambda\alpha]. \tag{66}$$

Combining (66) with (65) yields the solution for $\mathbb{E}\mu$:

$$\mathbb{E}\mu = \frac{\lambda(1-\gamma\beta)}{1-\gamma\beta\left(2-\beta\right) - \left[\lambda(1-\gamma\beta)\right]^2}\left[(1+\lambda)\Delta - \lambda\alpha\right],$$

and combining this solution with (59) yields the solution for μ^H:

$$\mu^H = \frac{1-\gamma\beta\left(2-\beta\right) + \lambda(1-\gamma\beta)}{1-\gamma\beta\left(2-\beta\right) - \left[\lambda(1-\gamma\beta)\right]^2}\left[(1+\lambda)\Delta - \lambda\alpha\right]. \tag{67}$$

The contradiction assumption therefore requires

$$\mu^H - \gamma\beta\mathbb{E}\mu = \frac{1-\gamma\beta\left(2-\beta\right) + \lambda\left(1-\gamma\beta\right) - \lambda\gamma\beta(1-\gamma\beta)}{1-\gamma\beta\left(2-\beta\right) - \left[\lambda(1-\gamma\beta)\right]^2}\left[(1+\lambda)\Delta - \lambda\alpha\right] > \Delta. \tag{68}$$

Recall that by Step 1e, $\lambda < 1$. Thus, the denominator above satisfies

$$1-\gamma\beta\left(2-\beta\right) - \left[\lambda(1-\gamma\beta)\right]^2 > 1-\gamma\beta\left(2-\beta\right) - (1-\gamma\beta)^2 \geq 0. \tag{69}$$

Since, by Assumption 1, $(1+\lambda)\Delta - \lambda\alpha < \Delta(1-\lambda)$, it follows that (68) requires

$$\left[1-\gamma\beta\left(2-\beta\right) + \lambda\left(1-\gamma\beta\right) - \lambda\gamma\beta(1-\gamma\beta)\right](1-\lambda) > 1-\gamma\beta\left(2-\beta\right) - \left[\lambda(1-\gamma\beta)\right]^2.$$

This inequality simplifies to

$$(1-\gamma\beta)^2 > 1-\gamma\beta\left(2-\beta\right),$$

which is a contradiction. The claim thus follows.

Step 1g. We show that the solution satisfies $\mu^L \geq -\Delta + \gamma\beta\mathbb{E}\mu$, which implies $\mu^H > -\Delta+\gamma\beta\mathbb{E}\mu$. Suppose by contradiction that $\mu^L < -\Delta+\gamma\beta\mathbb{E}\mu$. Combining the solution for $\mathbb{E}\mu$ in Step 1f with (58) yields the solution for μ^L:

$$\mu^L = -\frac{1-\gamma\beta\,(2-\beta)-\lambda(1-\gamma\beta)}{1-\gamma\beta\,(2-\beta)-[\lambda(1-\gamma\beta)]^2}\left[(1+\lambda)\Delta-\lambda\alpha\right].$$

The contradiction assumption therefore requires

$$\mu^L-\gamma\beta\mathbb{E}\mu = -\left\{\frac{1-\gamma\beta\,(2-\beta)-\lambda\,(1-\gamma\beta)+\lambda\gamma\beta(1-\gamma\beta)}{1-\gamma\beta\,(2-\beta)-[\lambda(1-\gamma\beta)]^2}\right\}\left[(1+\lambda)\Delta-\lambda\alpha\right] < -\Delta. \tag{70}$$

Note that the term in curly brackets is strictly between 0 and 1. Thus, (70) requires

$$(1+\lambda)\Delta-\lambda\alpha > \Delta, \tag{71}$$

which is a contradiction by Assumption 1. The claim thus follows.

Step 2. We verify that the solution to the relaxed problem satisfies the constraints of the original problem. The binding constraint (20) for $j=L$ implies

$$\Phi\left(\theta^*-s^L|0,\sigma^2\right)\overline{P} = \left(\alpha+s^L+\gamma\beta\mathbb{E}\mu\right)\left(\mu^H-\mu^L\right)-\frac{\left(\mu^H\right)^2-\left(\mu^L\right)^2}{2}. \tag{72}$$

Since $s^H > s^L$ and $\mu^H > \mu^L$, the right-hand side of (72) is strictly smaller than

$$\left(\alpha+s^H+\gamma\beta\mathbb{E}\mu\right)\left(\mu^H-\mu^L\right)-\frac{\left(\mu^H\right)^2-\left(\mu^L\right)^2}{2}. \tag{73}$$

Moreover, the left-hand side of (72) is strictly larger than

$$\Phi\left(\theta^*-s^H|0,\sigma^2\right)\overline{P}. \tag{74}$$

Therefore, (73) is strictly larger than (74), implying that constraint (20) for $j=H$ is satisfied.

To verify that constraint (21) for $j=L$ is satisfied, note that Steps 1f and 1g imply $\mu^L \in [s^L+\gamma\beta\mathbb{E}\mu, \alpha+s^L+\gamma\beta\mathbb{E}\mu]$. This implies that the low type's welfare in the optimal rule is no less than that under $\mu^L = s^L+\gamma\beta\mathbb{E}\mu$. Thus, (21) for $i=L$ is guaranteed to hold if

$$\overline{P} \geq \frac{\alpha^2}{2},$$

which is satisfied by Assumption 2.

Finally, we verify that constraint (21) for $i = H$ is also satisfied. Note that by constraint (20) for $i = H$ being satisfied, the high type's welfare in the optimal rule is no less than that achieved from mimicking the low type under $\mu^L = s^L + \gamma\beta\mathbb{E}\mu$. Thus, (21) for $i = H$ is guaranteed to hold if

$$\overline{P} \geq \frac{(2\Delta + \alpha)^2}{2},$$

which is satisfied by Assumption 1 and Assumption 2.

Step 3. We verify that a maximally-enforced hybrid threshold $\{\mu^*, \mu^{**}, \{\pi^*(\mu)\}_{\mu \in \mathbb{R}}\}$ implements the solution. Let $\mu^* = \mu^L$ and $\mu^{**} = \mu^H$. Construct the function $\pi^*(\mu)$ as described in (24), with $h(\mu)$ solving

$$\begin{aligned} \Phi\left(\mu - h\left(\mu\right)|s^L, \sigma^2\right)\overline{P} &= \left(\alpha + s^L + \gamma\beta\mathbb{E}\mu\right)\mu - \frac{\mu^2}{2} \\ &\quad - \left(\alpha + s^L + \gamma\beta\mathbb{E}\mu\right)\mu^L + \frac{(\mu^L)^2}{2}. \end{aligned} \tag{75}$$

The left-hand side of (75) is increasing in $\mu - h\left(\mu\right)$ and the right-hand side is increasing in μ. Note that $\lim_{\mu \downarrow \mu^*} h\left(\mu\right) = \infty$ and, by (72), $h\left(\mu^{**}\right) = \mu^{**} - \theta^*$. It follows that a solution for $h\left(\mu\right)$ exists and $\mu - h\left(\mu\right)$ is increasing in μ.

We verify that both central bank types choose their prescribed policies, $\mu^L = \mu^*$ and $\mu^H = \mu^{**}$, under this maximally-enforced hybrid threshold. By Step 2, neither type has incentives to deviate to $\mu > \mu^{**}$, as the best such deviation yields welfare weakly below that associated with setting $\mu = \alpha + s^j + \gamma\beta\mathbb{E}\mu$, for $j = L, H$, and is thus unattractive by (21). The low type has no incentive to deviate to $\mu = \mu^{**}$ by (20) for $j = L$, and this type has no incentive to deviate to $\mu < \mu^*$ either as he is better off by instead choosing $\mu^* < \alpha + s^L + \gamma\beta\mathbb{E}\mu$ and receiving no punishment. The high type has no incentive to deviate to $\mu \leq \mu^*$, as the best such deviation entails choosing $\mu^* < \alpha + s^H + \gamma\beta\mathbb{E}\mu$ which is unattractive by (20) for $j = H$. Therefore, it only remains to be shown that neither type has incentives to deviate to $\mu \in (\mu^*, \mu^{**})$. This follows immediately from (75) for the low type, as this equation ensures that the low type is indifferent between choosing μ^* and choosing any $\mu \in (\mu^*, \mu^{**})$. To show that the high type has no incentive to deviate to $\mu \in (\mu^*, \mu^{**})$, combine (72) and (75) to obtain:

$$\begin{aligned}\left(\Phi\left(\theta^*|s^L,\sigma^2\right) - \Phi\left(\mu - h\left(\mu\right)|s^L,\sigma^2\right)\right)\overline{P} &= \left(\alpha + s^L + \gamma\beta\mathbb{E}\pi\right)\mu^H - \frac{(\mu^H)^2}{2} \\ &\quad - \left(\alpha + s^L + \gamma\beta\mathbb{E}\pi\right)\mu + \frac{\mu^2}{2}.\end{aligned} \tag{76}$$

Since $s^H > s^L$ and $\mu^H > \mu$, the right-hand side of (76) is strictly smaller than

$$\left(\alpha + s^H + \gamma\beta\mathbb{E}\pi\right)\mu^H - \frac{(\mu^H)^2}{2} - \left(\alpha + s^H + \gamma\beta\mathbb{E}\pi\right)\mu + \frac{\mu^2}{2}. \tag{77}$$

Moreover, note that $\mu - h(\mu) < \theta^*$ for $\mu \in (\mu^*, \mu^{**})$, and $\theta^* < 0$ by Step 1e. Hence, the left-hand side of (76) is strictly larger than

$$\left(\Phi\left(\theta^*|s^H,\sigma^2\right) - \Phi\left(\mu - h\left(\mu\right)|s^H,\sigma^2\right)\right)\overline{P}. \tag{78}$$

Therefore, (77) is strictly larger than (78), implying that the high type has no incentive to deviate to $\mu \in (\mu^*, \mu^{**})$.

Online Appendix for
"Instrument-Based vs. Target-Based Rules"

by Marina Halac and Pierre Yared

B Proofs for Section 5

B.1 Proof of Proposition 5

B.1.1 Optimal Instrument-Based Rule

The program that solves for the optimal instrument-based rule under a continuum of types is analogous to that in (12) in Subsection 3.1 of the paper:

$$\max_{\{\mu(s),P(s)\}_{s\in\mathbb{R}}} \int_{-\infty}^{\infty} \left[U(\mu(s),\theta,\mathbb{E}\mu) - P(s)\right] \phi(\theta|s,\sigma^2)\phi(s|0,\Delta^2)d\theta ds \tag{79}$$

subject to, for all $s, s' \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} \left[\alpha\mu(s) + U(\mu(s),\theta,\mathbb{E}\mu) - P(s)\right] \phi(\theta|s,\sigma^2)d\theta \geq \int_{-\infty}^{\infty} \left[\alpha\mu(s') + U(\mu(s'),\theta,\mathbb{E}\mu) - P(s')\right] \phi(\theta|s,\sigma^2)d\theta, \tag{80}$$

$$\int_{-\infty}^{\infty} \left[\alpha\mu(s) + U(\mu(s),\theta,\mathbb{E}\mu) - P(s)\right] \phi(\theta|s,\sigma^2)d\theta \geq \int_{-\infty}^{\infty} \left[\alpha\mu^f(s,\mathbb{E}\mu) + U(\mu^f(s,\mathbb{E}\mu),\theta,\mathbb{E}\mu) - \overline{P}\right] \phi(\theta|s,\sigma^2)d\theta, \tag{81}$$

$$\mathbb{E}\mu = \int_{-\infty}^{\infty} \mu(s)\phi(s|0,\Delta^2)ds, \tag{82}$$

$$P(s) \in [0,\overline{P}]. \tag{83}$$

We begin by characterizing the optimal instrument-based rule for $\gamma = 0$. In this circumstance, $U(\mu(s),\theta,\mathbb{E}\mu)$ is independent of $\mathbb{E}\mu$ and (82) is a non-binding constraint.

An instrument-based rule specifying $\{\mu(s), P(s)\}_{s\in\mathbb{R}}$ is incentive compatible if this allocation satisfies (80)-(81), and it is incentive compatible and feasible, or incentive feasible for short, if the allocation satisfies (80)-(83). As mentioned in the paper, we assume that an optimal instrument-based rule is piecewise continuously differentiable. Additionally, if the program above admits multiple solutions that differ only on a countable set of types, we select the solution that maximizes social welfare for those types.

We proceed in four steps. Step 1 establishes some preliminary results that we use in the subsequent steps. Step 2 shows that any optimal instrument-based rule must prescribe bang-bang punishments. Step 3 shows that in any such rule, either no type is punished, or there exists an interior cutoff s^{**} such that only types above s^{**} receive the maximal punishment. Step 4 concludes the proof by characterizing the policy allocation and showing that such an interior cutoff s^{**} indeed exists in any optimal instrument-based rule.

Step 1. *We establish some preliminary results.*

The next lemma follows from standard arguments; see Fudenberg and Tirole (1991):

Lemma 2. $\{\mu(s), P(s)\}_{s\in\mathbb{R}}$ *satisfies the private information constraint* (80) *if and only if: (i)* $\mu(s)$ *is nondecreasing, and (ii) the following local private information constraints are satisfied:*

1. *At any point* s *at which* $\mu(\cdot)$*, and thus* $P(\cdot)$*, are differentiable,*

$$\mu'(s)(s+\alpha-\mu(s))-P'(s)=0.$$

2. *At any point* s *at which* $\mu(\cdot)$ *is not differentiable,*

$$\lim_{s'\uparrow s}\left\{(s+\alpha)\mu(s')-\frac{\mu(s')^2}{2}-P(s')\right\}=\lim_{s'\downarrow s}\left\{(s+\alpha)\mu(s')-\frac{\mu(s')^2}{2}-P(s')\right\}.$$

The private information constraints imply that the derivative of the central bank's welfare with respect to s is $\mu(s)$. Hence, in an incentive compatible rule, the welfare of type $s\in\mathbb{R}$ satisfies

$$(s+\alpha)\mu(s)-\frac{\mu(s)^2}{2}-P(s)=\lim_{\underline{s}\to-\infty}\left\{(\underline{s}+\alpha)\mu(\underline{s})-\frac{\mu(\underline{s})^2}{2}-P(\underline{s})+\int_{\underline{s}}^{s}\mu(\widetilde{s})d\widetilde{s}\right\}. \quad (84)$$

Following Amador, Werning, and Angeletos (2006), we can substitute (84) into the social welfare objective in (79) to rewrite this objective as

$$\lim_{\underline{s}\to-\infty}\left\{(\underline{s}+\alpha)\mu(\underline{s})-\frac{\mu(\underline{s})^2}{2}-P(\underline{s})+\int_{\underline{s}}^{\infty}\mu(s)Q(s)ds\right\}, \tag{85}$$

where

$$Q\left(s\right)\equiv 1-\Phi(s|0,\Delta^2)-\alpha\phi(s|0,\Delta^2).$$

Note that

$$Q'(s)=-\phi(s|0,\Delta^2)-\alpha\phi'(s|0,\Delta^2),$$

and thus $Q'(s)<0$ if $s<\widehat{s}\equiv\Delta^2/\alpha$ and $Q'(s)>0$ if $s>\widehat{s}$. (Observe that this property on $Q'(s)$ holds for some $\widehat{s}$ for any density function ϕ that is log concave.) We next describe two functions that we will use in our proofs.

Lemma 3. *Given $s^L\leq s^M\leq s^H$, define the functions*

$$\begin{aligned} B^L(s^L,s^M) &= \int_{s^L}^{s^M}\left(s-s^L-\alpha\right)\phi(s|0,\Delta^2)ds+\alpha\phi(s^M|0,\Delta^2)(s^M-s^L),\\ B^H(s^H,s^M) &= \int_{s^M}^{s^H}\left(s-s^H-\alpha\right)\phi(s|0,\Delta^2)ds+\alpha\phi(s^M|0,\Delta^2)(s^H-s^M). \end{aligned}$$

Then $B^L(s^L,s^M)>0$ if $Q'(s)<0$ for all $s\in(s^L,s^M)$, $B^L(s^L,s^M)<0$ if $Q'(s)>0$ for all $s\in(s^L,s^M)$, $B^H(s^H,s^M)>0$ if $Q'(s)>0$ for all $s\in(s^M,s^H)$, and $B^H(s^H,s^M)<0$ if $Q'(s)<0$ for all $s\in(s^M,s^H)$.

Proof. Consider the claims about $B^L(s^L,s^M)$. Note that $B^L(s,s^M)|_{s=s^M}=0$, and hence $B^L(s^L,s^M)=-\int_{s_L}^{s^M}\frac{dB^L(s,s^M)}{ds}ds$. Moreover,

$$\frac{dB^L(s,s^M)}{ds}=-\int_s^{s^M}\phi(\widetilde{s}|0,\Delta^2)d\widetilde{s}+\alpha\phi(s|0,\Delta^2)-\alpha\phi(s^M|0,\Delta^2),$$

and thus $\frac{dB^L(s,s^M)}{ds}|_{s=s^M}=0$. Therefore, $B^L(s^L,s^M)=\int_{s_L}^{s^M}\int_s^{s^M}\frac{d^2B^L(\widetilde{s},s^M)}{d\widetilde{s}^2}d\widetilde{s}ds$, where

$$\frac{d^2B^L(s,s^M)}{ds^2}=\phi(s|0,\Delta^2)+\alpha\phi'(s|0,\Delta^2).$$

Note that $\frac{d^2B^L(s,s^M)}{ds^2}>0$ if $Q'(s)<0$, $\frac{d^2B^L(s,s^M)}{dB^2}=0$ if $Q'(s)=0$, and $\frac{d^2B^L(s,s^M)}{ds^2}<0$ if $Q'(s)>0$. The claims about $B^L(s^L,s^M)$ follow.

The proof for the claims about $B^H(s^H,s^M)$ is analogous and thus omitted. □

Step 2. *We show that if* $\{\mu(s), P(s)\}_{s\in\mathbb{R}}$ *is an optimal instrument-based rule, then* $P(s) \in \{0, \overline{P}\}$ *for all* $s \in \mathbb{R}$.

Take any solution to the program in (79)-(83). We proceed in three sub-steps.

Step 2a. *We show* $P(s)$ *is left-continuous at each* $s \in \mathbb{R}$.

Suppose by contradiction that there exists s at which $P(s)$ is not left-continuous. Denote the left limit by $\{\mu(s^-), P(s^-)\} = \lim_{s'\uparrow s}\{\mu(s'), P(s')\}$. By Lemma 2,

$$(s+\alpha)\mu(s) - \frac{\mu(s)^2}{2} - (s+\alpha)\mu(s^-) + \frac{\mu(s^-)^2}{2} = P(s) - P(s^-).$$

Given $\alpha > 0$ and the fact that $\mu(s^-) < \mu(s)$ by Lemma 2, this implies

$$s\mu(s) - \frac{\mu(s)^2}{2} - s\mu(s^-) - \frac{\mu(s^-)^2}{2} < P(s) - P(s^-).$$

It follows that a perturbation that assigns $\{\mu(s^-), P(s^-)\}$ to type s is incentive feasible, strictly increases social welfare from type s, and does not affect social welfare from types other than s. Hence, $P(s)$ must be left-continuous at each $s \in \mathbb{R}$.

Step 2b. *We show* $P(s)$ *is a step function over any interval* $[s^L, s^H]$ *with* $P(s) \in (0, \overline{P})$.

By the private information constraints, $P(s)$ is piecewise continuously differentiable and nondecreasing. Suppose by contradiction that there is an interval $[s^L, s^H]$ over which $P(s)$ is continuously strictly increasing in s and satisfies $0 < P(s) < \overline{P}$. By Lemma 2, $\mu(s)$ must be continuously strictly increasing over the interval, and without loss we can take an interval over which $\mu(s)$ is continuously differentiable. Moreover, by the properties of the normal distribution, we can take either an interval above $\widehat{s}$ with $Q'(s) > 0$ for all $s \in [s^L, s^H]$ or an interval below $\widehat{s}$ with $Q'(s) < 0$ for all $s \in [s^L, s^H]$. We consider each possibility in turn.

Case 1: Suppose $Q'(s) < 0$ for all $s \in [s^L, s^H]$. We show that there exists an incentive feasible perturbation that rotates the increasing schedule $\mu(s)$ clockwise over $[s^L, s^H]$ and strictly increases social welfare. Define

$$\overline{\mu} = \frac{1}{(s^H - s^L)} \int_{s^L}^{s^H} \mu(s) ds.$$

For given $\tau \in [0, 1]$, let $\widetilde{\mu}(s, \tau)$ be the solution to

$$\widetilde{\mu}(s,\tau) = \tau\overline{\mu} + (1-\tau)\,\mu(s), \tag{86}$$

which clearly exists. Define $\widetilde{P}(s,\tau)$ as the solution to

$$(s+\alpha)\widetilde{\mu}(s,\tau)-\frac{\widetilde{\mu}(s,\tau)^2}{2}-\widetilde{P}(s,\tau) = (s^L+\alpha)\mu(s^L)-\frac{\mu(s^L)^2}{2}-P(s^L)+\int_{s^L}^{s}\widetilde{\mu}(\widetilde{s},\tau)d\widetilde{s}. \quad (87)$$

The original allocation corresponds to $\tau = 0$. We consider a perturbation where we increase τ marginally above zero if and only if $s \in [s^L, s^H]$. Note that differentiating (86) and (87) with respect to τ yields

$$\frac{d\widetilde{\mu}\,(s,\tau)}{d\tau} = \overline{\mu} - \mu(s), \quad (88)$$

$$\frac{d\widetilde{\mu}\,(s,\tau)}{d\tau}\left(s+\alpha-\widetilde{\mu}(s,\tau)\right) - \frac{d\widetilde{P}(s,\tau)}{d\tau} = \int_{s^L}^{s}\frac{d\widetilde{\mu}(\widetilde{s},\tau)}{d\tau}d\widetilde{s}. \quad (89)$$

Substituting (88) in (89) yields that for a type $s \in [s^L, s^H]$, the change in the central bank's welfare from a marginal increase in τ, starting from $\tau = 0$, is equal to

$$D\,(s) \equiv \int_{s^L}^{s}\left(\overline{\mu} - \mu(\widetilde{s})\right)d\widetilde{s}.$$

We begin by showing that the perturbation satisfies constraints (80)-(83). For the private information constraint (80), note that $D(s^L) = D(s^H) = 0$, so the perturbation leaves the welfare of types s^L and s^H (and that of types $s < s^L$ and $s > s^H$) unchanged. Using Lemma 2 and the representation in (84), it then follows from equation (87) and the fact that $\widetilde{\mu}(s,\tau)$ is nondecreasing in s that the perturbation satisfies constraint (80) for all s and any $\tau \in [0,1]$.

To prove that the perturbation satisfies the enforcement constraint (81), we show that the welfare of types $s \in [s^L, s^H]$ weakly rises when τ increases marginally. Since $D(s^L) = D(s^H) = 0$, it is sufficient to show that $D(s)$ is concave over (s^L, s^H) to prove that $D(s) \geq 0$ for all s in this interval. Indeed, we verify:

$$D'(s) = \overline{\mu} - \mu(s),$$
$$D''(s) = -\mu'(s) < 0.$$

Lastly, observe that constraint (83) is satisfied for $\tau > 0$ small enough. This follows from the fact that $P(s) \in (0, \overline{P})$ for $s \in [s^L, s^H]$ in the original allocation.

We next show that the perturbation strictly increases social welfare. Using the representation in (85), the change in social welfare from an increase in τ is equal to

$$\int_{s^L}^{s^H} \frac{d\widetilde{\mu}\,(s,\tau)}{d\tau} Q(s) ds.$$

Substituting with (88) and the expression for $Q(s)$ yields that this is equal to

$$\int_{s^L}^{s^H} (\overline{\mu} - \mu(s)) \left(1 - \Phi(s|0,\Delta^2) - \alpha\phi(s|0,\Delta^2)\right) ds.$$

This is an integral over the product of two terms. The first term is strictly decreasing in s since $\mu(s)$ is strictly increasing over $[s^L, s^H]$. The second term is also strictly decreasing in s; this follows from $Q'(s) < 0$ for all $s \in [s^L, s^H]$. Therefore, these two terms are positively correlated with one another, and thus the change in social welfare is strictly greater than

$$\int_{s^L}^{s^H} (\overline{\mu} - \mu(s))\, ds \int_{s^L}^{s^H} \left(1 - \Phi(s|0,\Delta^2) - \alpha\phi(s|0,\Delta^2)\right) ds,$$

which is equal to 0. It follows that the change in social welfare from the perturbation is strictly positive. Hence, if $P(s)$ is strictly interior and $Q'(s) < 0$ over a given interval, then $P(s)$ must be a step function over the interval.

Case 2: Suppose $Q'(s) > 0$ for all $s \in [s^L, s^H]$. Recall that $\mu(s)$ is continuously strictly increasing over $[s^L, s^H]$. We begin by showing that the enforcement constraint (81) cannot bind for all $s \in [s^L, s^H]$. Suppose by contradiction that it does. Using the representation of the central bank's welfare in (84), this implies

$$\int_{s}^{s^H} (\widetilde{s} + \alpha - \mu(\widetilde{s}))d\widetilde{s} = 0$$

for all $s \in [s^L, s^H]$, which requires $\{\mu(s), P(s)\} = \{s + \alpha, \overline{P}\}$ for all $s \in (s^L, s^H)$. However, this contradicts the assumption that $P(s) \in (0, \overline{P})$ for all $s \in [s^L, s^H]$. Hence, the enforcement constraint cannot bind for all types in the interval, and without loss we can take an interval with this constraint being slack for all $s \in [s^L, s^H]$.

We next show that there exists an incentive feasible perturbation that strictly increases social welfare. Specifically, consider drilling a hole around a type s^M within $[s^L, s^H]$ so that we marginally remove the allocation around this type. That is, type s^M can no longer choose $\{\mu(s^M), P(s^M)\}$ and is indifferent between jumping to the lower or upper limit of the hole. With some abuse of notation, denote the limits of the hole by s^L and s^H, where the perturbation marginally increases s^H from s^M. Since the enforcement constraint is slack for all $s \in [s^L, s^H]$, the perturbation is incentive feasible. The change in social welfare from the perturbation is equal to

$$\int_{s^M}^{s^H} \left(\mu'(s^H)(s-\mu(s^H)) - P'(s^H)\right)\phi(s|0,\Delta^2)ds$$
$$+\frac{ds^M}{ds^H}\left(s^M\mu(s^L) - \frac{\mu(s^L)^2}{2} - P(s^L) - s^M\mu(s^H) + \frac{\mu(s^H)^2}{2} + P(s^H)\right)\phi(s^M|0,\Delta^2).$$

Note that by the private information constraint for type s^H,

$$\mu'(s^H)(s^H + \alpha - \mu(s^H)) - P'(s^H) = 0, \tag{90}$$

and by indifference of type s^M,

$$(s^M+\alpha)\mu(s^L) - \frac{\mu(s^L)^2}{2} - P(s^L) = (s^M+\alpha)\mu(s^H) - \frac{\mu(s^H)^2}{2} - P(s^H). \tag{91}$$

Substituting with these expressions, the change in social welfare is equal to

$$\mu'(s^H)\int_{s^M}^{s^H}\left(s - s^H - \alpha\right)\phi(s|0,\Delta^2)ds + \frac{ds^M}{ds^H}\alpha\left(\mu(s^H) - \mu(s^L)\right)\phi(s^M|0,\Delta^2). \tag{92}$$

Differentiating (91) with respect to s^H and substituting with (90) yields

$$\frac{ds^M}{ds^H} = \mu'(s^H)\frac{(s^H - s^M)}{\mu(s^H) - \mu(s^L)}.$$

Substituting back into (92) and dividing by $\mu'(s^H) > 0$, we find that the change in social welfare takes the same sign as

$$B^H(s^H, s^M) = \int_{s^M}^{s^H}\left(s - s^H - \alpha\right)\phi(s|0,\Delta^2)ds + \alpha\phi(s^M|0,\Delta^2)(s^H - s^M).$$

Since $Q'(s) > 0$ for all $s \in [s^M, s^H]$, Lemma 3 implies $B^H(s^H, s^M) > 0$, and thus the perturbation strictly increases social welfare. Hence, if $P(s)$ is strictly interior and $Q'(s) > 0$ over a given interval, then $P(s)$ must be a step function over the interval.

Step 2c. *We show $P(s) \in \{0, \overline{P}\}$ for all $s \in \mathbb{R}$.*

Suppose by contradiction that $P(s) \in (0, \overline{P})$ for some s. By the previous steps and Lemma 2, s belongs to a stand-alone segment $(s^L, s^H]$, such that $\mu(s) = \mu$ and $P(s) = P$ for all $s \in (s^L, s^H]$, with $P \in (0, \overline{P})$ (by assumption), and $\mu(s)$ jumps at s^L and s^H.

We first show that the enforcement constraint must be slack for all $s \in (s^L, s^H)$. Express the enforcement constraint as the difference between the left-hand and right-hand sides of (81), so that this constraint must be weakly positive and it is equal to zero if it binds. By the private information constraints, the derivative of the enforcement constraint with respect to s is equal to $\mu(s) - \alpha - s$. Since $\mu(s)$ is constant over $(s^L, s^H]$, it follows that the enforcement constraint is strictly concave over the interval, and therefore slack for all $s \in (s^L, s^H)$.

We next show that there exists an incentive feasible perturbation that strictly increases social welfare. We consider perturbations that marginally change the constant policy μ and punishment P. As we describe next, the perturbation that we perform depends on the shape of the function $Q(s)$ over $(s^L, s^H]$:

Case 1: Suppose $\int_{s^L}^{s^H} Q(s^L)ds < \int_{s^L}^{s^H} Q(s)ds$. Consider a perturbation that marginally changes the constant policy by $d\mu > 0$ and increases P in order to keep type s^H equally well off. This means that $\frac{dP}{d\mu}$ is given by

$$s^H + \alpha - \mu - \frac{dP}{d\mu} = 0. \tag{93}$$

Note that for any arbitrarily small $d\mu > 0$, this perturbation makes the lowest types s in $(s^L, s^H]$, arbitrarily close to s^L, jump either to the allocation of type s^L or to their flexible allocation under maximal punishment $\{s+\alpha, \overline{P}\}$, where we let the perturbation introduce the latter. In the limit as $d\mu$ goes to zero, the change in social welfare due to the perturbation is thus equal to[37]

$$\begin{aligned}\int_{s^L}^{s^H} \left(s - \mu - \frac{dP}{d\mu}\right)\phi(s|0,\Delta^2)ds \\ + \frac{ds^L}{d\mu}\left(s^L\mu(s^L) - \frac{\mu(s^L)^2}{2} - P(s^L) - s^L\mu + \frac{\mu^2}{2} + P\right)\phi(s^L|0,\Delta^2),\end{aligned} \tag{94}$$

where the following indifference condition holds:

$$(s^L + \alpha)\mu - \frac{\mu^2}{2} - P = (s^L + \alpha)\mu(s^L) - \frac{\mu(s^L)^2}{2} - P(s^L).$$

[37]The arguments that follow are unchanged if $\{\mu(s^L), P(s^L)\}$ is replaced with $\{s^L + \alpha, \overline{P}\}$ for the cases where the enforcement constraint binds.

To verify that the perturbation is incentive feasible, note that the enforcement constraint is slack for all $s \in (s^L, s^H)$, P is strictly interior, and the welfare of types s^L and s^H remains unchanged with the perturbation. Hence, the perturbation is incentive feasible for $d\mu$ arbitrarily close to zero.

To verify that the perturbation strictly increases social welfare, substitute (93) and the indifference condition of type s^L into (94) to obtain:

$$\int_{s^L}^{s^H} \left(s - s^H - \alpha\right) \phi(s|0, \Delta^2) ds + \frac{ds^L}{d\mu} \alpha \left(\mu(s^L) - \mu\right) \phi(s^L|0, \Delta^2). \tag{95}$$

Differentiating the indifference condition of type s^L and substituting with (93) yields

$$\frac{ds^L}{d\mu} = \frac{s^H - s^L}{\mu - \mu(s^L)}.$$

Substituting back into (95), we find that the change in social welfare takes the same sign as

$$B^H(s^H, s^L) = \int_{s^L}^{s^H} \left(s - s^H - \alpha\right) \phi(s|0, \Delta^2) ds + \alpha \phi(s^L|0, \Delta^2)(s^H - s^L),$$

which can be rewritten as

$$B^H(s^H, s^L) = \int_{s^L}^{s^H} \int_{s^L}^{s} Q'(\widetilde{s}) d\widetilde{s} ds = \int_{s^L}^{s^H} (Q(s) - Q(s^L)) ds.$$

By the assumption that $\int_{s^L}^{s^H} Q(s^L) ds < \int_{s^L}^{s^H} Q(s) ds$, the above expression is strictly positive. The perturbation therefore strictly increases social welfare, yielding a contradiction.

Case 2: Suppose $\int_{s^L}^{s^H} Q(s^L) ds \geq \int_{s^L}^{s^H} Q(s) ds$. Since $Q'(s) \neq 0$ almost everywhere, there must exist $s^h \in (s^L, s^H]$ such that $\int_{s^L}^{s^h} Q(s^L) ds > \int_{s^L}^{s^h} Q(s) ds$. Then consider a perturbation where, for $s \in (s^L, s^h]$, we marginally change the constant policy by $d\mu < 0$ and decrease P in order to keep type s^h equally well off. This perturbation makes types arbitrarily close to s^L jump up to the allocation of the stand-alone segment. Arguments analogous to those in Case 1 above imply that the perturbation is incentive feasible. Moreover, following analogous steps as in that case yields that the implied change in social welfare takes the same sign as

$$-B^H(s^h, s^L) = -\int_{s^L}^{s^h} \left(s - s^h - \alpha\right) \phi(s|0, \Delta^2) ds - \alpha\phi(s^L|0, \Delta^2)(s^h - s^L),$$

which can be rewritten as

$$-B^H(s^h, s^L) = -\int_{s^L}^{s^h} \int_{s^L}^{s} Q'(\widetilde{s}) d\widetilde{s} ds = -\int_{s^L}^{s^h} (Q(s) - Q(s^L)) ds.$$

By the assumption that $\int_{s^L}^{s^h} Q(s^L) ds > \int_{s^L}^{s^h} Q(s) ds$, the above expression is strictly positive. The perturbation therefore strictly increases social welfare, yielding a contradiction.

Step 3. *We show that if $\{\mu(s), P(s)\}_{s\in\mathbb{R}}$ is an optimal instrument-based rule, then either $P(s) = 0$ for all $s \in \mathbb{R}$, or there exists some $s^{**} \in (\widehat{s}, \infty)$ such that $P(s) = 0$ if $s \leq s^{**}$ and $P(s) = \overline{P}$ if $s > s^{**}$.*

Take any solution to the program in (79)-(83). We proceed in two sub-steps.

Step 3a. *We show that if $P(s^{**}) = \overline{P}$ for some $s^{**} \in \mathbb{R}$, then $s^{**} \geq \widehat{s}$.*

By Step 2a, if $P(s^{**}) = \overline{P}$ for some s^{**}, then $P(s) = \overline{P}$ over an interval $(s^L, s^H]$ that contains s^{**}. Take the largest such interval. We establish that $s^L \geq \widehat{s}$. Suppose by contradiction that $s^L < \widehat{s}$ and take a subinterval $(s^L, s^h]$ below $\widehat{s}$. Note that the enforcement constraint requires $\mu(s) = s + \alpha$ for all $s \in (s^L, s^h]$. Then we can perform a perturbation that rotates the policy schedule clockwise over this interval, analogous to the perturbation used in Case 1 in Step 2b. By the arguments in that case, this perturbation is incentive feasible. In particular, note that since the perturbation weakly increases the welfare of all types $s \in (s^L, s^h]$ while simultaneously changing their policy away from $s+\alpha$, it follows that the perturbation must necessarily decrease $P(s)$ below $\overline{P}$. Moreover, by $Q'(s) < 0$ for all types $s \in (s^L, s^h]$ (by this interval being below $\widehat{s}$), the perturbation strictly increases social welfare, yielding a contradiction.

Step 3b. *We show that if $P(s^{**}) = \overline{P}$ for some $s^{**} \in \mathbb{R}$, then $P(s) = \overline{P}$ for all $s \geq s^{**}$.*

Suppose by contradiction that $P(s^{**}) = \overline{P}$ for $s^{**} \in (-\infty, \infty)$ and $P(s) < \overline{P}$ for some $s > s^{**}$. By Step 3a above, $s^{**} \geq \widehat{s}$. Moreover, by Step 2 (and the contradiction assumption), there exist $s^H > s^L \geq s^{**}$ such that $P(s) = 0$ for all $s \in (s^L, s^H]$.

We begin by establishing that $\mu(s) = \mu$ for all $s \in (s^L, s^H]$ and some μ. Suppose by contradiction that $\mu'(s) > 0$ at some $s' \in (s^L, s^H]$. Note that the private information constraint (80) (together with the constant punishment over $(s^L, s^H])$ implies $\mu(s) = s + \alpha$, and thus a slack enforcement constraint, in the neighborhood of such type s'. Then we can perform an incentive feasible perturbation that drills a hole in the $\mu(s)$ schedule in this neighborhood, as that described in Case 2 in Step 2b. By the arguments in that case, this perturbation strictly increases social welfare, yielding a contradiction.

We next show that a segment $(s^L, s^H]$ with $\mu(s) = \mu$ and $P(s) = 0$ for all $s \in (s^L, s^H]$ and $s^L \geq s^{**}$ cannot exist. Suppose by contradiction that it does. Take s^L to be the lowest point weakly above s^{**} at which P jumps, and take s^H to be the lowest point above s^L at which P jumps again. Note that $s^H < \infty$ must exist, since (81) cannot be satisfied for all $s > s^L$ with $\mu(s) = \mu$ and $P(s) = 0$ for all such s. Then $(s^L, s^H]$ is a stand-alone segment with constant policy μ and zero punishment. Note that by arguments analogous to those in Step 2c, the enforcement constraint must be slack for all $s \in (s^L, s^H)$. Moreover, observe that $\mu < s^H + \alpha$ must hold, since otherwise by Lemma 2 and the monotonicity of $\mu(s)$, (80) would be violated at s^H. It follows that we can perform an incentive feasible perturbation analogous to that used in Step 2c: for $\mu' = \mu + \varepsilon, \varepsilon > 0$ arbitrarily small, we increase μ marginally to μ' and set P slightly above 0 so as to keep type s^H's welfare under this allocation unchanged. Since $s^L \geq s^{**}$ implies $\int_{s^L}^{s^H} Q(s^L)ds < \int_{s^L}^{s^H} Q(s)ds$, this perturbation strictly increases social welfare, yielding a contradiction.

Step 4. *We characterize the optimal policy allocation and show that any optimal instrument-based rule specifies $s^{**} \in (\widehat{s}, \infty)$ as defined in Step 3.*

Take any solution to program (79)-(83). By Step 3, either $P(s) = 0$ for all $s \in \mathbb{R}$, or there exists some $s^{**} \in (\widehat{s}, \infty)$ such that $P(s) = 0$ for $s \leq s^{**}$ and $P(s) = \overline{P}$ for $s > s^{**}$. In the latter case, by the enforcement constraint (81), $\mu(s) = s + \alpha$ for all $s > s^{**}$, and since (81) holds with equality at s^{**}, this type's allocation satisfies

$$(s^{**} + \alpha)\mu(s^{**}) - \frac{\mu(s^{**})^2}{2} = \frac{(s^{**} + \alpha)^2}{2} - \overline{P}. \tag{96}$$

These results characterize the allocation for types $s \geq s^{**}$ when there exists an interior type s^{**} as defined above. We next proceed by characterizing the allocation that corresponds either to types $s < s^{**}$ in this case, or to all types in the case that such an interior type s^{**} does not exist. The final step of the proof establishes that the latter scenario does not arise, namely any optimal instrument-based rule specifies an interior type s^{**} such that $P(s) = 0$ for $s \leq s^{**}$ and $P(s) = \overline{P}$ for $s > s^{**}$.

Step 4a. *We show $\mu(s)$ is continuous over any interval $[s^L, s^H]$ such that $P(s) = 0$ for all $s \in [s^L, s^H]$.*

There are two cases to consider:

Case 1: Suppose by contradiction that $\mu(s)$ has a point of discontinuity below $\widehat{s}$. Note that if s^{**} as defined above exists, then the assumed point of discontinuity is strictly below s^{**}. The discontinuity requires that a type $s^M < \widehat{s}$ be indifferent between choosing $\lim_{s \uparrow s^M} \mu(s)$ and $\lim_{s \downarrow s^M} \mu(s) > \lim_{s \uparrow s^M} \mu(s)$. Note that given $P(s) = 0$ around this point, there must be a hole with types $s \in [s^L, s^M)$ bunched at $\mu(s^L) = s^L + \alpha$ and types $s \in (s^M, s^H]$ bunched at $\mu(s^H) = s^H + \alpha$, for some $s^L < s^M < s^H$. Now consider perturbing this rule by marginally increasing s^L, in an effort to slightly close the hole. This perturbation leaves the welfare of types strictly above s^M unchanged and is incentive feasible. The change in social welfare from the perturbation is equal to

$$\mu'(s^L) \int_{s^L}^{s^M} \left(s - \mu(s^L)\right) \phi(s|0, \Delta^2) ds$$
$$+ \frac{ds^M}{ds^L} \left(s^M \mu(s^L) - \frac{\mu(s^L)^2}{2} - s^M \mu(s^H) + \frac{\mu(s^H)^2}{2} \right) \phi(s^M|0, \Delta^2).$$

Note that $\mu(s^L) = s^L + \alpha$, $\mu(s^H) = s^H + \alpha$, and $\mu'(s^L) = 1$. Moreover, by indifference of type s^M, we have

$$(s^M + \alpha)\mu(s^L) - \frac{\mu(s^L)^2}{2} = (s^M + \alpha)\mu(s^H) - \frac{\mu(s^H)^2}{2}. \tag{97}$$

Substituting into the expression above yields that the change in social welfare is equal to

$$\int_{s^L}^{s^M} \left(s - s^L - \alpha\right) \phi(s|0, \Delta^2) ds + \frac{ds^M}{ds^L} \phi(s^M|0, \Delta^2) \alpha \left(\mu(s^H) - \mu(s^L)\right). \tag{98}$$

Note that differentiating the indifference condition (97) with respect to s^L (and substituting with $\mu(s^L) = s^L + \alpha$) yields

$$\frac{ds^M}{ds^L} = \frac{(s^M - s^L)}{\mu(s^H) - \mu(s^L)}.$$

Substituting this back into (98), we find that the change in social welfare is equal to

$$B^L(s^L, s^M) = \int_{s^L}^{s^M} \left(s - s^L - \alpha\right) \phi(s|0, \Delta^2) ds + \alpha\phi(s^M|0, \Delta^2)(s^M - s^L).$$

It follows from $s^M < \widehat{s}$ and Lemma 3 that $B^L(s^L, s^M) > 0$. Thus, the perturbation strictly increases social welfare, showing that $\mu(s)$ cannot jump at a point below $\widehat{s}$.

Case 2: Suppose by contradiction that $\mu(s)$ has a point of discontinuity above $\widehat{s}$. Note that if s^{**} as defined above exists, then the assumed point of discontinuity is strictly below s^{**}. By the same logic as in Step 3b, we can show that $\mu'(s) = 0$ over any continuous interval above $\widehat{s}$ over which $P(s) = 0$. It follows that there must exist a stand-alone segment $(s^L, s^H]$ with constant policy μ and zero punishment, satisfying $s^L \geq \widehat{s}$. However, using again the arguments in Step 3b, a perturbation that marginally increases μ and sets P slightly above 0 would then be incentive feasible and would strictly increase social welfare. Therefore, $\mu(s)$ cannot jump at a point above $\widehat{s}$ around which $P(s) = 0$.

Step 4b. *We show $\mu(s) \leq s + \alpha$ for all s for which $P(s) = 0$.*

Consider types $s \leq s^{**}$ when an interior point s^{**} as described above exists, or all types $s \in \mathbb{R}$ when such a point s^{**} does not exist. Step 4a above implies that the allocation for these types must be bounded discretion, with either a minimum policy level or a maximum policy level or both. Note that if a minimum policy level is prescribed, then there must exist some interior point $s^\star$ such that the allocation satisfies $\{\mu(s), P(s)\} = \{s^\star + \alpha, 0\}$ for all $s \leq s^\star$. However, such an allocation would violate the enforcement constraint (81) for s sufficiently low. Therefore, a minimum policy level is not enforceable and only a maximum policy level can be imposed, establishing the claim.

Step 4c. *We show $\mu(s) < s + \alpha$ for some s for which $P(s) = 0$. Moreover, there exists $s^{**} \in (\widehat{s}, \infty)$ as defined in Step 3.*

Suppose that an interior point s^{**} as described above exists. By Steps 4a and 4b, (96) must hold for $\mu(s^{**}) = s^* + \alpha$, where the value of s^* is unique given s^{**} and satisfies $s^{**} = s^* + \sqrt{2\overline{P}}$. In this circumstance, the optimal instrument-based rule is implemented with a strictly interior threshold $\mu^* = s^* + \alpha$, and the central bank's policy satisfies $\mu(s) < s + \alpha$ for all $s \in (s^*, s^{**})$.

We end the proof by showing that an interior point s^{**} as described above must indeed exist in any optimal instrument-based rule. Suppose by contradiction that this is not the case. Then by the steps above, there must be an optimal instrument-based rule prescribing $\{\mu(s), P(s)\} = \{s + \alpha, 0\}$ for all $s \in \mathbb{R}$. Using the representation in (85), social welfare under this rule is equal to

$$\lim_{\underline{s} \to -\infty} \left\{ \frac{(\underline{s} + \alpha)^2}{2} + \int_{\underline{s}}^{\infty} (s + \alpha) Q(s) ds \right\}. \tag{99}$$

Consider social welfare under a maximally enforced threshold $\mu^* = s^* + \alpha$:

$$\lim_{\underline{s} \to -\infty} \left\{ \frac{(\underline{s} + \alpha)^2}{2} + \int_{\underline{s}}^{\infty} (s + \alpha) Q(s) ds + \int_{s^*}^{s^{**}} [(s^* + \alpha) - (s + \alpha)] Q(s) ds \right\}, \tag{100}$$

where $s^{**} = s^* + \sqrt{2\overline{P}}$. The contradiction assumption requires that (99) weakly exceed (100) for all strictly interior s^*. That is, for all $s^* \in (-\infty, \infty)$ and $s^{**} = s^* + \sqrt{2\overline{P}}$, the following condition must hold:

$$\int_{s^*}^{s^{**}} (s^* - s) \, Q(s) ds \leq 0. \tag{101}$$

Note that $\lim_{s \to \infty} Q(s) = 0$ and $Q'(s) > 0$ for all $s > \widehat{s}$. Thus, setting $s^* \geq \widehat{s}$ yields $Q(s) < 0$ for all $s \in [s^*, s^{**}]$. This implies that the left-hand side of (101) is an integral over the product of two negative terms, and thus strictly positive, yielding a contradiction.

We have characterized the optimal instrument-based rule for $\gamma = 0$. We next show that the same proof applies to the case of $\gamma = 1$ after following a change of variable. Specifically, let $\widehat{\mu}(s) \equiv \mu(s) - \beta \mathbb{E} \mu(s)$ and $\widehat{\mu}^f(s) \equiv s + \alpha$. Then given $\gamma = 1$, (79)-(83)

can be rewritten as

$$\max_{\{\widehat{\mu}(s),P(s)\}_{s\in\mathbb{R}}} \int_{-\infty}^{\infty} \left[U(\widehat{\mu}(s),\theta,0) - P(s)\right] \phi(\theta|s,\sigma^2)\phi(s|0,\Delta^2)d\theta ds$$

subject to, for all $s, s' \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} \left[\alpha\widehat{\mu}(s) + U(\widehat{\mu}(s),\theta,0) - P(s)\right] \phi(\theta|s,\sigma^2)d\theta$$
$$\geq \int_{-\infty}^{\infty} \left[\alpha\widehat{\mu}(s') + U(\widehat{\mu}(s'),\theta,0) - P(s')\right] \phi(\theta|s,\sigma^2)d\theta,$$

$$\int_{-\infty}^{\infty} \left[\alpha\widehat{\mu}(s) + U(\widehat{\mu}(s),\theta,0) - P(s)\right] \phi(\theta|s,\sigma^2)d\theta$$
$$\geq \int_{-\infty}^{\infty} \left[\alpha\widehat{\mu}^f(s) + U(\widehat{\mu}^f(s),\theta,0) - \overline{P}\right] \phi(\theta|s,\sigma^2)d\theta,$$

$$P(s) \in [0,\overline{P}].$$

This program is the same as (79)-(83) under $\gamma = 0$. Hence, our previous arguments apply directly with respect to the schedule $\widehat{\mu}(s)$, proving in turn the results for the schedule $\mu(s)$ under $\gamma = 1$.

B.1.2 Optimal Target-Based Rule

The characterization of the optimal target-based rule under a continuum of types follows the same steps as in our baseline model. Letting $\mu(s)$ denote the policy of type s, condition (18) characterizes the value of $\delta = \mu(s) - s$. The optimal target-based rule can then be represented as the solution to (34), where it is clear that this solution is independent of the distribution of types. As such, the arguments in the proof of Proposition 2 apply directly to this setting with a continuum of types.

B.2 Proof of Proposition 6

The proof of Proposition 5 shows that, within each rule class, the optimal rule is the same under $\gamma = 0$ and $\gamma = 1$. We thus proceed to prove Proposition 6 for the case of $\gamma = 0$, and our same arguments apply to the case of $\gamma = 1$.

The optimal target-based rule under a continuum of signals, and social welfare under this rule, are identical to those in our baseline model. Therefore, the same arguments as in the proof of Lemma 1 (in the proof of Proposition 3) apply. Those arguments imply that social welfare is strictly decreasing in σ under the optimal target-based rule. In fact, the proof of Lemma 1 shows that the derivative of social welfare with respect to σ is strictly lower than $-\sigma$ under this rule.

To compare with the optimal instrument-based rule, we prove the following lemma:

Lemma 4. *Suppose $\gamma = 0$. Consider changing σ while keeping $Var(\theta)$ unchanged. The change in social welfare from a marginal increase in σ is strictly higher than $-\sigma$ under the optimal instrument-based rule.*

Proof. Social welfare under the optimal instrument-based rule can be written as

$$-\frac{\sigma^2}{2} + \int_{-\infty}^{\infty} \left[-\frac{(s-\mu(s))^2}{2} - P(s) \right] \phi(s|0,\Delta^2)ds. \tag{102}$$

Substituting with the structure of the optimal rule yields

$$-\frac{\sigma^2}{2} + \left\{ \int_{-\infty}^{s^*} -\frac{\alpha^2}{2}\phi(s|0,\Delta^2)ds + \int_{s^*}^{s^{**}} \left[-\frac{(s-s^*-\alpha)^2}{2} \right] \phi(s|0,\Delta^2)ds \right.$$
$$\left. + \int_{s^{**}}^{\infty} \left(-\frac{\alpha^2}{2} - \overline{P} \right) \phi(s|0,\Delta^2)ds \right\}, \tag{103}$$

where $s^{**} = s^* + \sqrt{2\overline{P}}$. Since Δ declines as σ rises, it is sufficient to show that the term in curly brackets in (103) is decreasing in Δ. To evaluate the derivative of this term, define $\widetilde{s} = s/\Delta$, with $\widetilde{s}^* = s^*/\Delta$ and $\widetilde{s}^{**} = s^{**}/\Delta$, where $\widetilde{s}^{**} = \widetilde{s}^* + \frac{1}{\Delta}\sqrt{2\overline{P}}$. Using integration by substitution, the term in curly brackets in (103) can be written as

$$\int_{-\infty}^{\widetilde{s}^*} -\frac{\alpha^2}{2}\phi(\widetilde{s}|0,1)d\widetilde{s} + \int_{\widetilde{s}^*}^{\widetilde{s}^{**}} \left[-\frac{\left(\Delta(\widetilde{s}-\widetilde{s}^{**}) + \sqrt{2\overline{P}} - \alpha\right)^2}{2} \right] \phi(\widetilde{s}|0,1)d\widetilde{s}$$
$$+ \int_{\widetilde{s}^{**}}^{\infty} \left(-\frac{\alpha^2}{2} - \overline{P} \right) \phi(\widetilde{s}|0,1)d\widetilde{s}. \tag{104}$$

Since the optimal instrument-based rule selects values for $\widetilde{s}^*$ and $\widetilde{s}^{**}$ to maximize (104), this rule necessarily satisfies the following first-order condition:

$$\int_{\widetilde{s}^*}^{\widetilde{s}^{**}} \left(\Delta(\widetilde{s}-\widetilde{s}^{**}) + \sqrt{2\overline{P}} - \alpha \right) \phi(\widetilde{s}|0,1)ds = -\alpha(\widetilde{s}^{**} - \widetilde{s}^*)\phi(\widetilde{s}^{**}|0,1) < 0. \tag{105}$$

The derivative of (103) with respect to Δ, taking into account the Envelope condition, is thus equal to

$$-\int_{\widetilde{s}^{*}}^{\widetilde{s}^{**}}(\widetilde{s}-\widetilde{s}^{**})\Big(\Delta(\widetilde{s}-\widetilde{s}^{**})+\sqrt{2\overline{P}}-\alpha\Big)\phi(\widetilde{s}|0,1)ds. \tag{106}$$

Both terms in the integral are increasing in $\widetilde{s}$, which means that (106) takes the same sign as

$$-\int_{\widetilde{s}^{*}}^{\widetilde{s}^{**}}(\widetilde{s}-\widetilde{s}^{**})\phi(\widetilde{s}|0,1)ds\int_{\widetilde{s}^{*}}^{\widetilde{s}^{**}}\Big(\Delta(\widetilde{s}-\widetilde{s}^{**})+\sqrt{2\overline{P}}-\alpha\Big)\phi(\widetilde{s}|0,1)ds. \tag{107}$$

The first integral in (107) is negative since $\widetilde{s}<\widetilde{s}^{**}$ for all $\widetilde{s}\in(\widetilde{s}^{*},\widetilde{s}^{**})$. The second integral is also negative by (105). Therefore, (107) is strictly negative. It follows that the second term in (103) is decreasing in Δ, establishing the claim. □

We now proceed with the proof of the proposition. By Lemma 4, when σ increases, social welfare under the optimal instrument-based rule declines by less than social welfare under the optimal target-based rule. To prove the claim in the proposition, it is thus sufficient to show that, among these two rule classes, a target-based rule is optimal at one extreme, for $\sigma\to 0$, whereas an instrument-based rule is optimal at the other extreme, for $\sigma\to\sqrt{Var(\theta)}$. This is what we prove next.

Take first $\sigma\to 0$ and thus $\Delta\to\sqrt{Var(\theta)}$. The same arguments as those used in the proof of Proposition 3 imply that social welfare under the optimal target-based rule approaches 0 in this limit. Using the representation in (102), social welfare under the optimal instrument-based rule approaches

$$\lim_{\Delta\to\sqrt{Var(\theta)}}\left\{\int_{-\infty}^{\infty}\left[-\frac{(s-\mu(s))^2}{2}-P(s)\right]\phi(s|0,\Delta^2)ds\right\}.$$

Note that this expression can only exceed 0 if $\mu(s)=s$ and $P(s)=0$ for all $s\in\mathbb{R}$, but such an allocation would violate the private information constraint (80). Therefore, this expression must be strictly lower than 0. It follows that the optimal target-based rule dominates the optimal instrument-based rule for $\sigma\to 0$.

Take next $\sigma\to\sqrt{Var(\theta)}$ and thus $\Delta\to 0$. The same arguments as those used in the proof of Proposition 3 imply that social welfare under the optimal target-based rule approaches a value strictly lower than $-\frac{Var(\theta)}{2}$ in this limit. Social welfare under the optimal instrument-based rule approaches

$$-\frac{Var(\theta)}{2} + \lim_{\Delta\to 0}\left\{\int_{-\infty}^{\infty}\left[-\frac{(s-\mu(s))^2}{2} - P(s)\right]\phi(s|0,\Delta^2)ds\right\}. \tag{108}$$

In the limit as $\Delta \to 0$, $\phi(s|0,\Delta^2)$ corresponds to a Dirac's delta function, with cumulative distribution function $\Phi(s|0,\Delta^2) = 0$ if $s < 0$ and $\Phi(s|0,\Delta^2) = 1$ if $s \geq 0$. Consider the limiting social welfare under an instrument-based threshold specifying $s^* = -\alpha < 0$ and $s^{**} = -\alpha + \sqrt{2\overline{P}}$, where $s^{**} > 0$ by Assumption 2. Under this rule, $\mu(s) = 0$ if $s \in [s^*, s^{**}]$ and $\mu(s) = s + \alpha$ otherwise, and $P(s) = 0$ if $s \leq s^{**}$ and $P(s) = \overline{P}$ otherwise. Therefore, (108) under this rule becomes equal to $-\frac{Var(\theta)}{2}$. Since social welfare under the optimal instrument-based rule must weakly exceed this value, it follows that the optimal instrument-based rule dominates the optimal target-based rule for $\sigma \to \sqrt{Var(\theta)}$.