WORKING PAPER SERIES

Escapist Policy Rules

James Bullard
and
In-Koo Choo

Working Paper 2002-002C
http://research.stlouisfed.org/wp/2002/2002-002.pdf

January 2002
Revised January 2005

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
411 Locust Street
St. Louis, MO 63102

Escapist Policy Rules

James Bullard[*]

Federal Reserve Bank of St. Louis

and

In-Koo Cho[†]

University of Illinois

This version:[‡] *6 January 2005*

We study a simple, microfounded macroeconomic system in which the monetary authority employs a Taylor-type policy rule. We analyze situations in which the self-confirming equilibrium is unique and learnable according to Bullard and Mitra (2002). We explore the prospects for the use of 'large deviation' theory in this context, as employed by Sargent (1999), Williams (2004), and Cho, Williams, and Sargent (2002). We show that our system can sometimes depart from the self-confirming equilibrium towards a non-equilibrium outcome characterized by persistently low nominal interest rates and persistently low inflation. Thus we generate events that have some of the properties of "liquidity traps" observed in the data, even though the policymaker remains committed to a Taylor-type policy rule which otherwise has desirable stabilization properties.

Key Words: Learning, monetary policy rules, escape dynamics.

[*] Research Department, Federal Reserve Bank of St. Louis, 411 Locust Street, St. Louis, MO 63102 USA. Telephone: (314) 444-8576. Email: bullard@stls.frb.org. Fax: (314) 444-8731. Any views expressed are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

† Department of Economics, 330 Wohlers Hall, University of Illinois at Urbana-Champaign, 1206 S. Sixth Street, Champaign, IL 61820 USA. Email: inkoocho@uiuc.edu. Financial support from the National Science Foundation is gratefully acknowledged.

‡ This paper was originally prepared for a workshop on "Learning and Model Misspecification," in Cleveland, Ohio. We thank the Federal Reserve Bank of Cleveland for sponsoring this event, and John Carlson for organizing it. We also thank discussants Stephanie Schmitt-Grohe, Bob Tetlow, Leopold von Thadden, two anonymous refer-

1. INTRODUCTION

1.1. Overview

In the recent literature on monetary policy in microfounded models, there has been a great deal of discussion concerning nominal interest rate feedback rules as a guide for policymakers.[1] Generally speaking, the advice emanating from this literature has been that central banks could achieve near-optimal macroeconomic outcomes if they committed to a Taylor-type policy rule that has a certain property. This property is what Woodford (2001) dubs the "Taylor principle"—the rule must call for the central bank to change nominal interest rates sufficiently aggressively in response to inflation developments in the economy.[2] The conventional wisdom is thus that a monetary authority implementing a rule obeying the Taylor principle would probably do quite well with respect to minimizing fluctuations in inflation and real output.

In this paper we explore the robustness of this conventional wisdom to small departures from the extreme rationality assumptions that underlie it. We want to take a first step in this literature toward understanding how certain types of seemingly minor misspecifications along with agent learning might combine to change the global dynamics of the economy in unexpected ways. To pose this question, we start with a workhorse model from this literature, in order to remain generally consistent with other authors in this area. We endow the policymakers with a commitment to a policy rule obeying the Taylor principle. Thus, in a conventional analysis, we would conclude that this monetary policy was close to the optimal one. We alter the economy relative to this benchmark, in part by allowing the agents to use a slightly misspecified model of the economy, and in part by allowing the agents to learn over time instead of endowing them with rational expectations. In this altered economy, we find that the Taylor-type policy

ees, and seminar participants at the Texas Monetary Economics Conference, Midwest Macroeconomics, Princeton University, Federal Reserve Macro System Committe, and the Society for Economic Dynamics for helpful comments. In addition, we thank the organizers and participants at the conference "Expectations, Learning, and Monetary Policy," sponsored by the Deutsche Bundesbank and the Center for Financial Studies, in Frankfurt, Germany, for support and insightful comments.

[1] For a sample of the recent work, see Taylor (1993), the volumes edited by Taylor (1999) and King and Plosser (1999), and the survey by Clarida, Gali and Gertler (1999).

[2] This is sometimes also referred to as an "active" policy rule.

rule can still be quite successful, as the economy can remain in a small neighborhood of the unique self-confirming equilibrium for long periods of time. But we also provide conditions under which the system may abruptly *escape* from a neighborhood of that equilibrium towards a persistent low-nominal-interest-rate, non-equilibrium outcome. This escape outcome has some of the "liquidity trap" characteristics present in Japanese data from the 1990s, which we now describe.

1.2. The specter of Japan

Our themes in this paper are theoretical, and we do not intend to confront the Japanese data directly. But we do think that the Japanese experience can help to motivate the type of phenomenon we analyze below. During the middle-to-late 1980s, the Japanese economy was widely admired in the business press and among academics. It had grown rapidly for many years, and seemed to threaten U.S. world economic leadership. But Japanese success faded in the 1990s as the economy became mired in a cycle of poor performance. One of the features of the 1990s Japanese experience was a sharp decline in short-term nominal interest rates. Figure 1 shows annualized three-month unregulated time deposit rates in Japan from 1990 through 2000. These rates have remained below one percent per annum since 1995, after beginning the decade near four percent. The low nominal interest rates have been associated with low inflation rates. Consumer prices were rising at a rate of 3 to 4 percent per year in Japan at the beginning of the 1990s, but the inflation rate has fallen to between ± 1 percent since 1995, when measured as a percent increase from the previous year (the exception is 1997, when it rose to about two percent). Real performance has been poor during the 1990s, especially when compared to earlier decades.[3]

Policymaking at the Bank of Japan is sometimes suspected of causing the change of fortunes. To critics, if the Bank of Japan had somehow behaved differently than it did, the 1990s Japanese experience might have been avoided. A difficult aspect of the critics' view is that the Bank of

[3]Summers (1991) has argued that low nominal interest rates leave the economy more vulnerable to negative shocks, since monetary policymakers targeting nominal interest rates can do little when an adverse shock is realized.

Figure 1

Nominal interest rates and inflation in Japan 1990-2000

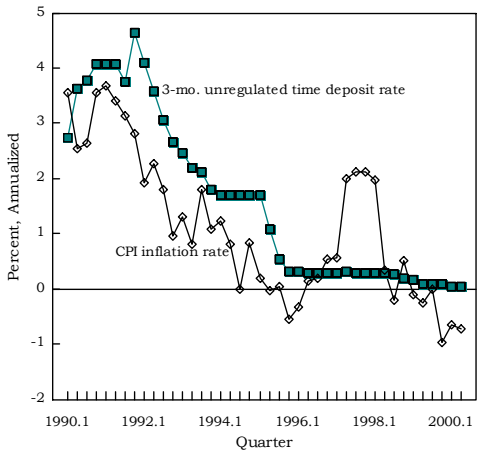

FIG. 1. Short-term nominal interest rates in Japan during the 1990s fell dramatically, in tandem with the CPI inflation rate.

Japan did not appear to behave very differently during the 1990s than it had during the earlier, more successful periods for the economy. If the Bank of Japan's policy rule was the right one during the successful periods, why was essentially the same policy rule the wrong one during the 1990s?

This paper has a lot to say about this type of question. We view Japanese monetary policymakers as using essentially the same monetary policy rule during the 1990s as they did during the earlier portions of the postwar era.[4] In fact, the policymakers in our model follow a Taylor-type policy rule from which they never deviate. This is of course an extreme assumption, but it is also a strength of our analysis, because it makes it clear that our dynamics are not generated by a change in the policy rule. Instead, the system *endogenously deviates* from the targeted equilibrium toward the liquidity trap outcome. As we will discuss in detail in the main text of the

[4]We could also think in terms of U.S. data. In the U.S., short-term nominal interest rates fell precipitously during the Great Depression and remained near zero for many years. See Wheelock (1991) for a discussion of the hypothesis that monetary policymakers at the time did not alter their operating procedure in any fundamental way.

paper, under certain circumstances a self-reinforcing process can begin in the neighborhood of the self-confirming equilibrium, propelling the system far from the targeted outcome. From the perspective of the agents in the model, this turn of events would be puzzling, since the policy rule is unchanged and produced quite good performance for a long period of time.[5]

1.3. What we do

We begin with a standard New Keynesian model as described by Woodford (1999, 2003) and Clarida, Gali, and Gertler (1999). We introduce learning into this economy, following the analysis of Bullard and Mitra (2002). We restrict attention to situations under which the targeted equilibrium of the central bank would be both determinate and learnable under their analysis. We then look for circumstances under which the stability under learning might break down, and cause the system to visit a low nominal interest rate, low inflation outcome, like the ones displayed in Figure 1.[6] We use 'large deviation' theory as employed by Sargent (1999) and Cho, Williams, and Sargent (2002) to generate these departures, or "escapes." We spend much of the paper documenting that the escape dynamics depend on three factors. These factors are (1) A certain misspecification on the part of the private sector regarding the actions of the policy authorities, (2) Feedback from the beliefs of the private sector to the actions of the policy authority, and (3) A learning rule that reflects the private sector's doubt about the accuracy of their specification. We think all three factors are plausibly at work in actual economies.

1.4. Recent related literature

Benhabib, Schmitt-Grohe and Uribe (2001) argued that the interaction between an active Taylor-type rule, a Fisher relation, and a zero bound on

[5]We are describing unintentionally low nominal interest rates as undesirable. Low nominal interest rates have sometimes been associated with poor economic performance in actual economies like Japan. In many contexts in monetary theory, however, low nominal interest rates are welfare-improving (see, for instance, Woodford [1990b]). We think of our problem as one where, for reasons exogenous to the model, the nominal interest rate and the inflation rate associated with the self-confirming equilibrium are socially optimal, and the goal of the government is to cause these values to come about. The large deviation from this equilibrium is then inadvertent and unwanted.

[6]Another way to put our primary question is to ask, "What assumptions are necessary to generate escape dynamics in this popular environment?"

nominal interest rates helps explain liquidity trap outcomes through the creation of a low inflation steady state. Our explanation is quite different from theirs, because we focus on a model with a single steady state and generate large deviations from that unique stable equilibrium point. For analyses of learning in environments more directly related to Benhabib, et al., (2001), see McCallum (2003), Eusepi (2003), and Evans and Honkapohja (2003). These latter authors discuss the circumstances under which the Benhabib, et al. (2001) liquidity trap equilibrium might be learnable.

The analysis here has the private sector learning and the central bank following a stipulated policy. For analyses in which the roles are reversed, see, for instance, Sargent (1999) and Wieland (2000).

2. ENVIRONMENT

2.1. A baseline economy

We study a model economy based on Woodford (1999, 2003). The model has been derived from microfoundations in Woodford and Rotemberg (1998) and elsewhere, and has become a workhorse model in the literature on monetary policy rules. We think the ideas we illustrate using this model would be equally applicable in related frameworks.

In our economy, economic time series are being generated in a manner subtly different from the world that the private sector agents perceive. However, the self-confirming equilibrium of the model has the private sector's perceptions verified by actual events, so that they do not discover the nature of their misspecified view of the economy. In order to build this type of model, we first show how the time series being generated depend on the perceptions of the agents, and then how the agents' perceptions differ from this reality. The actual evolution of the economy then depends on the interaction between these two dynamics.

Woodford's (1999, 2003) framework consists of two equations which are log-linear approximations to the first-order conditions for household and firm maximization problems in his economy. The households have a standard intertemporal optimization problem which yields a consumption Euler equation given by equation (1). The monopolistically competitive firms face frictions in setting nominal prices, and their profit maximization conditions

yield equation (2). Woodford's framework represents a simplified linearization about a steady state expressed in terms of the level of output z, the inflation rate π, and the nominal interest rate r:

$$z_t^d = \tilde{E}_t z_{t+1}^d - \sigma^{-1}\left(r_t^d - \tilde{E}_t \pi_{t+1}^d\right) + w_t \tag{1}$$

$$\pi_t^d = \kappa z_t^d + \beta \tilde{E}_t \pi_{t+1}^d \tag{2}$$

where

$$w_t = \alpha w_{t-1} + \epsilon_t, \tag{3}$$

and η_t and ϵ_t are Gaussian white noise terms. We let $z_t^d = z_t - \bar{z}_t$, $\pi_t^d = \pi_t - \bar{\pi}_t$, and $r_t^d = r_t - \bar{r}_t$, so that all variables are expressed as deviations from target or long-run values at time t denoted by $\bar{z}_t$, $\bar{\pi}_t$ and $\bar{r}_t$. We let $\tilde{E}_t$ be a (possibly nonrational) expectations operator representing the private sector's views of the future. The parameters σ, relating to the elasticity of intertemporal substitution of the representative household, κ, relating to the degree of price stickiness in the economy, and β, the common household discount factor, are all fixed and positive. We think of equations (1) and (2) as describing the optimizing behavior of the private sector, given their expectations, in Woodford's (1999) framework.

Rotemberg and Woodford (1998) argue that the coefficients σ^{-1}, β, and κ are invariant to the policy rule chosen by the monetary authorities for the determination of r_t, and they supplement these equations with various forms of Taylor-type policy rules to close the model. We follow their procedure. We use a Taylor-type policy rule

$$r_t^d = \phi_\pi \pi_t^d + \eta_t. \tag{4}$$

with

$$\phi_\pi > 1 \tag{5}$$

fixed where η_t is white noise, representing an unexpected shock to the nominal interest rate. The fact that $\phi_\pi > 1$ means that this policy rule is "active" in the nomenclature of the literature. We will sometimes refer to ϕ_π as the "degree of hawkishness" in the policy rule, because it describes

how aggressively the policy authority reacts to deviations of inflation from target. Also, since the coefficient ϕ_π is fixed, and the functional form is also fixed, the policymaker is committed to the use of the active Taylor-type policy rule. This is an important feature of our model which we will come back to throughout the paper.[7] Equation (4) may be rewritten in the form

$$r_t = \phi_{0,t} + \phi_\pi \pi_t + \eta_t \tag{6}$$

where $\phi_{0,t} \equiv (1 - \phi_\pi)\,\bar{\pi}_t + \rho$, and

$$\rho \equiv \beta^{-1} - 1 \tag{7}$$

is the fixed, long-run real rate of interest.

We supplement this model with a description of how the long-run or target values $\bar{z}_t$, $\bar{\pi}_t$, or $\bar{r}_t$ evolve over time. We think of the long-run level of output $\bar{z}$ as a constant, and we think of the long-run nominal interest rate as determined by a Fisher relation. Thus we have

$$\bar{z}_t = \bar{z} \quad \forall t, \tag{8}$$

and

$$\bar{r}_t = \bar{\pi}_t + \rho. \tag{9}$$

It therefore remains to describe how the government goes about setting its inflation target $\bar{\pi}_t$.

One of our key assumptions is that we view the government as indifferent to the exact target level of inflation within any reasonable bounds.[8] Because of this, the monetary authority is willing to acquiesce to a target level of inflation which is expected by the private sector, so long as the private sector expects some level that can be put under the rubric of "low inflation." We interpret policymakers in our model as having the view that they do not want to spend time potentially destabilizing the economy by trying to convince the public that the target is, say, 1.75 percent when the

[7] We do not impose an explicit lower bound on nominal interest rates, but we do ensure that such a bound is never violated in our simulations.

[8] Actual central banks often announce target ranges, for instance, so that they might be thought of as indifferent to exactly what inflation rate is achieved within the range.

public thinks it is 2.25 percent. Our policymakers are indifferent between two such targets, and so, as a tie-breaking rule among potential targets, they simply set their target to the one that the private sector expects. And indeed, nearly all of the time in our model, the inflation rate will remain close to the target that the private sector expects. We can express this assumption simply as

$$\bar{\pi}_t = \pi_t^\star \tag{10}$$

where $\pi_t^\star$ is the private sector's perceived inflation target.

If the government adopted a fixed target for inflation, then the escape dynamics we describe in the remainder of the paper could not occur. We need the "center of gravity" of the system to move slightly with incoming shocks. This will be a key aspect of our generation of escape dynamics in this framework. But—and this is quite important—the model with a fixed inflation target and the model with the moving target described here are observationally equivalent at the self-confirming equilibrium.[9] Thus policymakers could argue, as many actually do, that a precise statement of a numerical inflation target is not necessary to achieve satisfactory stabilization performance. They would be right, most of the time. But the fact that the inflation target moves slightly with incoming shocks opens the door to the possibility of escape dynamics, as we will demonstrate.

We stress that the government in our model is not trying to outwit the public. They behave mechanically. They are committed to using an active Taylor-type policy rule. They are also committed to producing the long-run level of inflation the public expects to get.

2.2. Private sector perceptions

The private sector agents in our model observe new data on output, inflation, and nominal interest rates in each period. They are endowed with a *perceived law of motion* for the economy, which is given by

$$z_t^d = c_{11} w_t + c_{12} \eta_t \tag{11}$$

[9]In addition, under some plausible conditions, there would be no escape from the self-confirming equilibrium and hence the two models would always be observationally equivalent. This is the case if private sector agents use recursive least squares algorithms, as described below.

and

$$\pi_t^d = c_{21} w_t + c_{22} \eta_t. \tag{12}$$

The set of coefficients $c = (c_{ij})_{i,j=1}^2$ represent the beliefs of the private sector about how output and inflation deviate from their respective targets. This perceived model is a good one because it corresponds exactly to the minimal state variable rational expectations equilibrium of this economy.

In endowing the private sector agents with the correct model of the equilibrium law of motion for the economy, up to the coefficients c, we are following Evans and Honkapohja (2001) and other authors in the learning literature. We are giving the agents a lot of information.[10] The assumption is very favorable to the agents being able to learn the rational expectations equilibrium. If the agents cannot learn the equilibrium under this very favorable assumption, then it is called into question whether such an equilibrium could be stable under learning in an actual economy. The equilibrium we study will indeed turn out to be learnable in the sense defined by Evans and Honkapohja (2001). Thus the "large deviation" dynamics we isolate are all the more remarkable.[11]

The private sector agents assume that a Fisher relation holds so that their perception of the nominal interest rate, $r_t^\star$, is

$$r_t^\star = \rho + \pi_t^\star. \tag{13}$$

The private sector agents also correctly assume that

$$\bar{z}_t = \bar{z} \quad \forall t \geq 1. \tag{14}$$

By fixing the mean of the real sector, we intentionally make any deviation from an equilibrium more difficult.

Our second key assumption concerns the nature of the private sector's beliefs concerning monetary policy. We use this belief to generate feedback

[10] Including knowledge of the shocks w_t and η_t. But not as much information as under rational expectations.

[11] We could, of course, study systems where the private sector agents do not have so much information about the economy. The agents could use a misspecified model, for instance, or they could be allowed only partial observation of information. But our idea is to show that large deviation dynamics can occur even under the Evans-Honkapohja, "minimal deviation from rational expectations" assumption for the agents' perceived law of motion.

between the perceptions of the private sector and the policy choices of the government. This feedback will be critical in generating escape dynamics. Our key assumption is given by equation (19) below, which depicts a linear relationship between $\pi_t^\star$ and an estimated, or perceived value of ϕ_π, governed by coefficients δ_0 and δ_1. In the next two paragraphs we motivate equation (19) as a natural outcome of a private sector belief in a convex Taylor rule.

We assume the private sector believes that the monetary authority uses a convex Taylor-type rule described by

$$r_t = \psi\left(\pi_t\right), \tag{15}$$

where $\psi' > 0$, $\psi'' > 0$, and $\psi\left(\cdot\right)$ is invertible. The convexity of $\psi\left(\cdot\right)$ implies that monetary policy responds more aggressively to inflation when inflation is higher, and less aggressively when inflation is lower. The precise form of the function $\psi\left(\cdot\right)$ describing beliefs is not necessary for our analysis. Given the perceptions (11) and (12), the private sector agents only need a conjecture for their perceived inflation target $\pi_t^\star$ in order to compute π_t^d.

To obtain $\pi_t^\star$, the private sector agents take the derivative of (15),

$$\frac{dr_t}{d\pi_t} = \psi'\left(\pi_t\right), \tag{16}$$

and then invert this equation to obtain

$$\pi_t = \left(\psi'\right)^{-1}\left(\frac{dr_t}{d\pi_t}\right). \tag{17}$$

The private sector agents correctly conjecture from (12) that actual inflation π_t on the left hand side of (17) is distributed about the inflation target $\pi_t^\star$. On the right hand side, an estimate of $dr_t/d\pi_t$ can be found by computing the coefficients in the simple auxiliary[12] regression

$$r_t = \hat{\phi}_{0,t} + \hat{\phi}_{\pi,t}\pi_t + \xi_t, \tag{18}$$

[12]It is auxiliary to the regression defined by equations (11) and (12), as we discuss below.

where ξ_t is the regression residual, and using $\hat{\phi}_{\pi,t}$ as the estimate of $dr_t/d\pi_t$. Using these facts, a proxy for the inflation target based on equation (17) is

$$\pi_t^\star = \delta_0 + \delta_1 \hat{\phi}_{\pi,t}, \tag{19}$$

where the right hand side is a linear approximation of $\left(\psi'\right)^{-1}\left(\hat{\phi}_{\pi,t}\right)$. Our key assumption is then simply (19), which keeps the entire model completely linear.

Once the private sector sets $\pi_t^\star$, $r_t^\star$ is determined by equation (13). Then, agents can calculate $\pi_t^d = \pi_t - \pi_t^\star$ from observed π_t to solve (11) and (12). We now turn to the question of how the actual environment interacts with the perceptions of the agents to generate a stationary equilibrium.

3. EQUILIBRIUM ANALYSIS

3.1. Self-confirming equilibrium

The private sector's model can be parameterized by $c = (c_{ij})_{i,j=1}^2$ and $\phi = (\phi_0, \phi_\pi)$. To survive a long series of observed data, (c, ϕ) must be consistent with the observations statistically. This consistency will determine the "equilibrium" model of the private agents. While this sort of consistency between the subjective beliefs of the decision maker and the observed data is one of the two pillars of rational expectations, our equilibrium concept differs from rational expectations equilibrium in the sense that we do not presume the class of models represented by (11), (12), (18), and (19) contains the true model. In this sense, we admit that the model of the private agents is misspecified. For this reason, we call our equilibrium concept *self-confirming* equilibrium.

DEFINITION 3.1. The pair (c, ϕ) is a self-confirming equilibrium if the distribution of (π_t, r_t, z_t) generated by (1), (2), (4), (8), (9) and (10) conditioned on (c, ϕ) is equal to the conditional distribution of (π_t, r_t, z_t) calculated from the private agent's model through (11), (12), (18), and (19).

Self-confirming equilibrium is milder than rational expectations equilibrium because the agents do not need to know the actual model entertained by the government. Still, the private agents have to know the equilibrium

distribution of (π_t, r_t, z_t) in order to calculate the equilibrium value of (c, ϕ). We shall relax this requirement later, when we examine the learning model.

We rearrange equations (1) and (2) as

$$\begin{bmatrix} z_t^d \\ \pi_t^d \end{bmatrix} = \begin{bmatrix} \frac{\sigma}{\kappa\phi_\pi+\sigma} & \frac{1-\beta\phi_\pi}{\kappa\phi_\pi+\sigma} \\ \frac{\kappa\sigma}{\kappa\phi_\pi+\sigma} & \frac{\kappa+\beta\sigma}{\kappa\phi_\pi+\sigma} \end{bmatrix} \begin{bmatrix} \tilde{E}_t z_{t+1}^d \\ \tilde{E}_t \pi_{t+1}^d \end{bmatrix} + \begin{bmatrix} \frac{\sigma}{\kappa\phi_\pi+\sigma} & \frac{1}{\kappa\phi_\pi+\sigma} \\ \frac{\kappa\sigma}{\kappa\phi_\pi+\sigma} & \frac{\kappa}{\kappa\phi_\pi+\sigma} \end{bmatrix} \begin{bmatrix} w_t \\ \eta_t \end{bmatrix} \quad (20)$$

or more compactly as

$$\begin{bmatrix} z_t^d \\ \pi_t^d \end{bmatrix} = \mathcal{B}\tilde{E}_t y_{t+1}^d + \mathcal{D} \begin{bmatrix} w_t \\ \eta_t \end{bmatrix}. \quad (21)$$

Since the private agent's model is confined to those which can be represented in the form of (11) and (12),

$$\tilde{E}_t \pi_{t+1}^d = c_{21}\alpha w_t \quad (22)$$

and

$$\tilde{E}_t z_{t+1}^d = c_{11}\alpha w_t. \quad (23)$$

These expectations can be substituted into equation (21) to obtain the actual values for z_t and π_t

$$\begin{bmatrix} z_t^d \\ \pi_t^d \end{bmatrix} = \mathcal{B} \begin{bmatrix} c_{11}\alpha w_t \\ c_{21}\alpha w_t \end{bmatrix} + \mathcal{D} \begin{bmatrix} w_t \\ \eta_t \end{bmatrix}, \quad (24)$$

or equivalently,

$$\begin{aligned} \begin{bmatrix} z_t^d \\ \pi_t^d \end{bmatrix} &= \mathcal{B} \begin{bmatrix} c_{11}\alpha & 0 \\ c_{21}\alpha & 0 \end{bmatrix} \begin{bmatrix} w_t \\ \eta_t \end{bmatrix} + \mathcal{D} \begin{bmatrix} w_t \\ \eta_t \end{bmatrix} \\ &= \left(\mathcal{B} \begin{bmatrix} c_{11}\alpha & 0 \\ c_{21}\alpha & 0 \end{bmatrix} + \mathcal{D} \right) \begin{bmatrix} w_t \\ \eta_t \end{bmatrix}. \end{aligned} \quad (25)$$

By replacing the left hand side by (11) and (12) and arranging terms, we have

$$\left(\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} - \mathcal{B} \begin{bmatrix} c_{11}\alpha & 0 \\ c_{21}\alpha & 0 \end{bmatrix} - \mathcal{D} \right) \begin{bmatrix} w_t \\ \eta_t \end{bmatrix} = 0 \quad (26)$$

which must hold for any value of (w_t, η_t) in equilibrium. Thus, the equilibrium value of $c = (c_{ij})_{i,j=1}^2$ obtains by solving

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} - \mathcal{B} \begin{bmatrix} c_{11}\alpha & 0 \\ c_{21}\alpha & 0 \end{bmatrix} - \mathcal{D} = 0 \quad (27)$$

as in Bullard and Mitra (2002). Let c^e be the equilibrium value of c.

In a self-confirming equilibrium, the slope $\hat{\phi}_\pi$ of the auxiliary regression

$$r_t = \hat{\phi}_0 + \hat{\phi}_\pi \pi_t \tag{28}$$

must satisfy

$$\hat{\phi}_\pi = \phi_\pi. \tag{29}$$

Once ϕ is determined, the equilibrium target inflation rate can be calculated according to (19) and (10). Then, the equilibrium target nominal interest rate is given by the Fisher equation, and therefore, all endogenous variables are determined in equilibrium.

Because the private sector correctly identifies ϕ_π according to equation (29), the self-confirming equilibrium outcome is observationally equivalent to the model in which the government is committed to the monetary policy rule (4) with a fixed inflation target.

3.2. Learnability

3.2.1. Decreasing gain algorithms

In a self-confirming equilibrium, the private sector agents have to know precisely the equilibrium distribution of π_t and r_t, which is too demanding to be a descriptive model of an economic agent. Instead, let us assume that the private sector agents recursively estimate $c = (c_{ij})_{i,j=1}^2$ and $\phi = (\phi_0, \phi_\pi)$. To differentiate the equilibrium value from the estimated value, we add a "hat" to the corresponding variable to denote the estimate, and by time subscript t, we mean the estimate based on the information available at the beginning of time t. Assuming that the private sector agents choose the estimator to minimize the forecasting error, the estimators for $c = (c_{ij})_{i,j=1}^2$ and $\phi = (\phi_0, \phi_\pi)$ evolve according to the following recursive least squares formulae:

$$\begin{bmatrix} \hat{c}_{11,t+1} \\ \hat{c}_{12,t+1} \end{bmatrix} = \begin{bmatrix} \hat{c}_{11,t} \\ \hat{c}_{12,t} \end{bmatrix} + a_t \Sigma_{w\eta,t}^{-1} \begin{bmatrix} w_t \\ \eta_t \end{bmatrix} \left[z_t^d - \hat{c}_{11,t} w_t - \hat{c}_{12,t} \eta_t \right], \tag{30}$$

$$\begin{bmatrix} \hat{c}_{21,t+1} \\ \hat{c}_{22,t+1} \end{bmatrix} = \begin{bmatrix} \hat{c}_{21,t} \\ \hat{c}_{22,t} \end{bmatrix} + a_t \Sigma_{w\eta,t}^{-1} \begin{bmatrix} w_t \\ \eta_t \end{bmatrix} \left[\pi_t^d - \hat{c}_{21,t} w_t - \hat{c}_{22,t} \eta_t \right], \tag{31}$$

$$\begin{bmatrix} \hat{\phi}_{0,t+1} \\ \hat{\phi}_{\pi,t+1} \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{0,t} \\ \hat{\phi}_{\pi,t} \end{bmatrix} + a_t \Sigma_{\pi,t}^{-1} \begin{bmatrix} 1 \\ \pi_t \end{bmatrix} \left(r_t - \hat{\phi}_{0,t} - \hat{\phi}_{\pi,t} \pi_t \right), \tag{32}$$

$$\Sigma_{w\eta,t+1} = \Sigma_{w\eta,t} + a_t \left(\begin{bmatrix} w_t^2 & w_t \eta_t \\ w_t \eta_t & \eta_t^2 \end{bmatrix} - \Sigma_{w\eta,t} \right), \tag{33}$$

and

$$\Sigma_{\pi,t+1} = \Sigma_{\pi,t} + a_t \left(\begin{pmatrix} 1 & \pi_t \\ \pi_t & \pi_t^2 \end{pmatrix} - \Sigma_{\pi,t} \right), \tag{34}$$

where $a_t > 0$ is the gain sequence, which is set as

$$a_t = \frac{1}{t} \tag{35}$$

for the recursive least squares learning algorithm.

An important question is whether and when the agent can learn the self-confirming equilibrium through the recursive least squares learning algorithm.

Definition 3.2. (Evans and Honkapohja (2001)) A self-confirming equilibrium (c^e, ϕ^e) is *learnable* if there is a $\mu > 0$ such that, if $(\hat{c}_0, \hat{\phi}_0)$ is in a μ-neighborhood of (c^e, ϕ^e), then

$$(\hat{c}_t, \hat{\phi}_t) \to (c^e, \phi^e) \tag{36}$$

with probability 1 where $(\hat{c}_t, \hat{\phi}_t)$ is generated by the least squares learning algorithm described above.

In order to show that the least squares learning algorithm converges, we borrow the machinery developed for stochastic approximation.[13] To simplify notation, let

$$x_t = \begin{bmatrix} \hat{c}_t \\ \hat{\phi}_t \\ \text{col}(\Sigma_{w\eta,t}) \\ \text{col}(\Sigma_{\pi,t}) \end{bmatrix} \tag{37}$$

and

$$v_t = \begin{bmatrix} w_t \\ \eta_t \end{bmatrix} \tag{38}$$

[13] See, for example, Kushner and Yin (1997).

and write the recursive formula (30), (31), (32), (33) and (34) compactly as

$$x_{t+1} = x_t + a_t Q(x_t, v_t). \tag{39}$$

Remark 3.1. It is a convention to contain x_t in a compact set, because the decision maker can easily identify that the estimator is out of the reasonable range if it becomes too large or too small. The most common method is to use the projection facility to contain x_t in a compact set.[14] Let $x_t = (x_{1,t}, \ldots, x_{\ell,t}) \in \Re^\ell$. Define

$$\Lambda = \prod_{k=1}^{\ell} [\underline{x}_k, \bar{x}_k] \tag{40}$$

where we choose $\underline{x}_k$ and $\bar{x}_k$ so that x^e is contained in the interior of Λ and along the boundary of Λ, the gradient induced by equation (44) below is pointing to the interior of Λ. That is, if $x_t \in \Lambda$ but $x_{t+1} \notin \Lambda$ according to (39), then x_{t+1} is "projected" back into some point in the interior of Λ. Thus, we have to adjust x_{t+1} according to the projection facility. Let

$$x_{t+1} = \lambda(x_t + a_t Q(x_t, v_t)) \tag{41}$$

be the "adjusted" learning algorithm by incorporating the projection facility λ. Although the selection of Λ is arbitrary and requires some knowledge about the location of x^e, we can usually choose Λ sufficiently large to include all "reasonable" values of x_t in practice. The role of the projection facility is only to ensure that x_t is contained in a compact set. Thus, in order to simplify notation, we shall drop the projection facility λ for the rest of the paper from the recursive formula, and instead, assume that there is a compact set Λ such that

$$x_t \in \Lambda \qquad \forall t \geq 1. \tag{42}$$

The first step is to extract the deterministic dynamics that is a reasonable approximation of the stochastic dynamics. Define

$$\bar{Q}(x) = \lim_{T \to \infty} \frac{1}{T} E\left[\sum_{t=1}^{T} Q(x_t, v_t)\right]. \tag{43}$$

[14] See, for example, Marcet and Sargent (1989) and Woodford (1990a).

By the mean dynamics, we mean the ordinary differential equation (ODE)

$$\dot{x} = \bar{Q}(x) \tag{44}$$

which is often called *the associated ODE*. Since the dynamics of the individual components are crucial for later analysis, it is useful to write down the associated ODE for each component:

$$\begin{bmatrix} \dot{\hat{\phi}}_0 \\ \dot{\hat{\phi}}_\pi \end{bmatrix} = \Sigma^{-1} \times \begin{bmatrix} \rho - \hat{\phi}_0 + (1 - \hat{\phi}_\pi)(\gamma_0 + \gamma_1 \hat{\phi}_\pi) \\ (\phi_\pi - \hat{\phi}_\pi)\sigma_\pi^2 + (\gamma_0 + \gamma_1 \phi_\pi)\left(\rho - \hat{\phi}_0 + (1 - \hat{\phi}_\pi)(\gamma_0 + \gamma_1 \hat{\phi}_\pi)\right) \end{bmatrix} \tag{45}$$

$$\dot{\Sigma} = \begin{bmatrix} 1 & \gamma_0 + \gamma_1 \hat{\phi}_\pi \\ \gamma_0 + \gamma_1 \hat{\phi}_\pi & (\gamma_0 + \gamma_1 \hat{\phi}_\pi)^2 + \sigma_\pi^2 \end{bmatrix} - \Sigma \tag{46}$$

where

$$\sigma_\pi^2 = \frac{\sigma_\epsilon^2}{1 - \alpha^2}. \tag{47}$$

The self-confirming equilibrium is the outcome that causes the right hand side of the ODE to vanish:

$$\hat{\phi}_\pi = \phi_\pi, \tag{48}$$

$$\hat{\phi}_0 = \rho + (1 - \phi_\pi)(\gamma_0 + \gamma_1 \phi_\pi), \tag{49}$$

and

$$\Sigma = \begin{bmatrix} 1 & \gamma_0 + \gamma_1 \hat{\phi}_\pi \\ \gamma_0 + \gamma_1 \hat{\phi}_\pi & (\gamma_0 + \gamma_1 \hat{\phi}_\pi)^2 + \sigma_\pi^2 \end{bmatrix}. \tag{50}$$

This proves that *if the learning process converges*, the private sector agents learn the true attitude of the government toward inflation, ϕ_π.

Kushner and Yin (1997) present the general conditions under which the stochastic recursive algorithms converge to the stable points of the associated ODE. We state the key result of Kushner and Yin (1997) adapted for our model:

THEOREM 3.1. *Suppose that the following conditions are satisfied: (1)* $a_t > 0$, $a_t \to 0$, $\sum_{t=1}^{T} a_t \to \infty$ *as* $T \to \infty$ *and* $\sum_{t=1}^{\infty} a_t^2 < \infty$, *(2)* v_t

is a martingale difference with bounded second moments, (3) Q in (39) is Lipschitz continuous, (4) The associated ODE (44) has a stable point x^e with a basin of attraction, (5) There is a compact set Λ such that $x_t \in \Lambda$ infinitely many times with probability 1. Then, for any initial condition $x_0 \in \Lambda$ for (39), $x_t \to x^e$ with probability 1.

It is straightforward to verify every condition for Theorem 3.1. From the right hand side of (44), we can calculate the stationary point of (44) where

$$\hat{\phi}_{\pi,t} = \phi_\pi \tag{51}$$

must hold. That is, in a self-confirming equilibrium, the private sector agents correctly infer the degree of the government's hawkishness toward inflation. Following Bullard and Mitra (2002), we can verify that the stationary point of the associated ODE (i.e., the self-confirming equilibrium) is stable if

$$\phi_\pi > 1. \tag{52}$$

Hence, as long as the government is committed to an active Taylor-type rule, the least squares learning algorithm converges to the self-confirming equilibrium. Furthermore, by following Bullard and Mitra (2002), we can also show that if

$$\phi_\pi < 1, \tag{53}$$

then the self-confirming equilibrium is not stable. From Evans and Honkapohja (2001) and Bullard and Mitra (2002), we know that $x_t \to x^e$ with probability 0 if $\phi_\pi < 1$. We conclude that if the private sector agents estimate (c, ϕ) according to the least squares learning algorithm, then the gain sequence a_t for the recursive algorithm satisfies the conditions of Theorem 3.1 and $(\hat{c}_t, \hat{\phi}_t)$ converges to the self-confirming equilibrium with probability 1.

The stability of the self-confirming equilibrium demonstrates that the observational equivalence between the policy rule (4) with a fixed inflation target and the monetary policy rule with a time-varying target is quite robust. In the self-confirming equilibrium, one cannot reject the hypothesis that the government is using an active Taylor-type rule with a fixed inflation

target. And, there is no possibility of escape so long as the private sector agents use recursive least squares where $a_t = 1/t$.

3.2.2. Fixed gain algorithms

However, in the least squares learning algorithm characterized by the gain sequence $a_t = 1/t$, the private sector agents presume that the underlying economy is stationary so that the data observed a long time ago is just as useful as the most recently observed data, and consequently, they assign equal weight to all past data. But if the private sector agents are a little suspicious about the stationarity of the underlying economy, the least squares learning algorithm is no longer a sensible way of estimating (c, ϕ).[15] In particular, if the agent observes that the estimated hawkishness of the government toward inflation is fluctuating, the stationarity of the economy and the commitment of the government to respond aggressively to inflation is called into question.

Under this hypothesis, a sensible learning algorithm assigns a larger weight to more recently observed data. One simple way of implementing this idea is to set

$$a_t = a > 0 \tag{54}$$

for a small positive constant a, so that the private sector agents can discount the influence of past observations at a geometric rate. In contrast to the least squares learning algorithm in which a_t is decreasing, we call the learning algorithm with $a_t = a$ a *fixed gain algorithm.*

One cannot apply Theorem 3.1 to show the convergence to the self-confirming equilibrium with probability 1 for a fixed gain algorithm, because one of the conditions of the theorem regarding a_t is violated. Yet, we can still prove that the invariance distribution converges to the self-confirming equilibrium in a weaker sense.

THEOREM 3.2. *(Benveniste, Metivier and Priouret (1990)) For each $a > 0$, the recursive learning algorithm has an invariance distribution of $(\hat{c}_t, \hat{\phi}_t)$. This invariance distribution converges weakly to the self-confirming equilibrium which is the stable point of the associated ODE.*

[15] For more discussion of constant gain algorithms and their macroeconomic applications, see Sargent (1999) and Orphanides and Williams (2003).

For a sufficiently small $a > 0$, $(\hat{c}_t, \hat{\phi}_t)$ must be distributed around the stable point of the associated ODE. Thus, many of the properties found in the least squares learning algorithm are carried over to the fixed gain algorithm.

Surprisingly, with a_t fixed to a small value $a > 0$, the dynamics of $(\hat{c}_t, \hat{\phi}_t)$ reveals rare but recurrent escapes from the self-confirming outcome (the stable solution of the associated ODE). In particular, $\hat{\phi}_{\pi,t}$ escapes from the stable point $\phi_\pi > 1$, and moves toward 1, before returning to a neighborhood of $\phi_\pi > 1$. Recall that the government is committed to a fixed value $\phi_\pi > 1$, which means that its attitude toward inflation remains equally hawkish at all times throughout the entire episode. But because the *perceived* hawkishness of the government's attitude toward inflation is fluctuating over time due to the shocks in the system, the target inflation rate is also fluctuating. As a result, as we will see, the inflation rate and the nominal interest rate may stay away from the stable point of the associated ODE (the self-confirming equilibrium) for an extended period.

We begin our analysis of this phenomenon with a quantitative illustration.

3.3. A quantitative illustration

The main qualitative feature of our simulation—that the system eventually displays a large deviation from the self-confirming equilibrium—is quite robust across parameter choices. But for purposes of illustration, we used the following parameter values. For the structural parameters, we took the calibrated values from Woodford (1999), $\sigma = 0.157$, $\kappa = 0.024$, and we set $\beta = .9975$. This means that the annualized real interest rate, $\rho = \beta^{-1} - 1$, is one percent in this example. We set $\delta_0 = -\rho$. In the stochastic processes, we set $\alpha = .9$, $\sigma_\epsilon = .00372$, and $\sigma_\eta = .002$. This represents a high degree of serial correlation and a low level of noise in the system relative to Woodford (1999). This is mainly so that the noise does not interfere with our observation of the escape dynamics. We keep the constant gain factor small by setting $a = .005$. We set $\delta_1 = 1/500$, a low value that shows how mild the inflation target dependence on private sector beliefs can be. This leaves only the coefficient in the government's Taylor-type rule to be

set. We want to choose a value that is consistent with both determinacy and learnability in the Bullard and Mitra (2002) analysis. This requires roughly that $\phi_\pi > 1$ in this model. Of course, we want to analyze an active Taylor-type rule as well, which also means $\phi_\pi > 1$. Accordingly, we set $\phi_\pi = 5$. In conjunction with $\delta_1 = 1/500$, this parameter choice means that the government's target inflation rate at the self-confirming equilibrium is 2.997 percent. The target nominal interest rate is then 4.0 percent for this example.

The escape outcome will turn out to be a situation where the perceived hawkishness of monetary policy, $\hat{\phi}_{\pi,t}$, is consistent with a Fisher relation instead of with a Taylor-type rule. In the Fisher relation, nominal interest rates move one-for-one with inflation, so that $\hat{\phi}_{\pi,t} = 1$, far more passive than the actual value of $\phi_\pi = 5$. The escape outcome is therefore characterized by values of -20 basis points for the inflation rate, and 80 basis points for the nominal interest rate.

Figure 2 shows the nominal interest rate dynamics for this example, in a simulation of 3,500 periods initialized at the self-confirming equilibrium. The system remains in a neighborhood of the self-confirming equilibrium for about 3,000 periods before an abrupt escape to the low nominal interest rate, low inflation outcome occurs.[16] The low nominal interest rate outcome is not a self-confirming equilibrium, because the private sector holds the belief $\hat{\phi}_\pi = 1$ when in fact the government's policy is unchanged and has $\phi_\pi = 5$. Therefore, even though it is not apparent from the figure, the system does *not* remain at the escape outcome indefinitely. Instead, the private sector gradually begins to discover that its estimate of monetary hawkishness is too low, and they begin to revise their estimate away from one and toward the actual value ϕ_π. This sends the system on a climb back toward the self-confirming equilibrium.

Figure 2 displays an abrupt escape from the self-confirming equilibrium, in the context of 3,500 observations of the nominal interest rate. In Figure 3, the dynamics of the nominal interest rate and the inflation rate are

[16]This figure makes it seem like an escape would hardly ever occur. However, somewhat larger values of the gain a cause escapes to occur much more frequently. We show this particular example to establish that even if the system remains in the vicinity of the self-confirming equilibrium for a very long time, an escape eventually occurs.

Figure 2. An abrupt departure

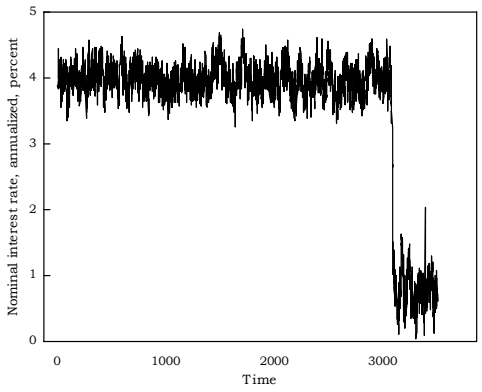

FIG. 2. A large deviation from the self-confirming equilibrium nominal interest rate. The system maintains interest rates in a neighborhood of 4.0 percent for many periods, but eventually the system departs to a low nominal interest rate outcome.

shown near the date of the escape—to obtain this figure, we selected 100 observations on the nominal interest rate as well as inflation for the period near the escape depicted in Figure 2. Figure 3 shows that the escape dynamics are abrupt, but not unrealistically so when compared to the actual Japanese data in Figure 1. The nominal interest rate falls from about 4 percent to just below 1 percent over a period of 16 quarters or so. The inflation rate also falls during this period. The inflation rate is relatively smooth in the figure because we do not have an inflation-specific shock in the model.

Despite the fact that the self-confirming equilibrium is globally stable under the learning algorithm, the model admits recurrent epochs of low inflation and low nominal interest rates. Our main interest is to understand the dynamics away from the stable self-confirming equilibrium from an analytical perspective. Therefore, we now turn away from the numerical example toward a proof of Proposition 3.1 below, which states the conditions under which escape dynamics occur. We stress that, under the assumptions we have made, escape dynamics are no accident but instead a probability one event.

Figure 3. Escape dynamics

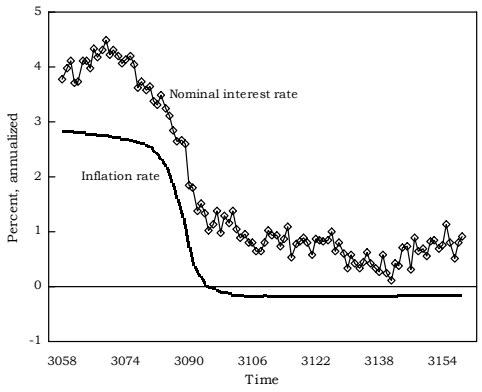

FIG. 3. A closer look at the escape dynamics. The nominal interest rate falls sharply over a period of say, 20 quarters, close to the observed timing in Japan documented in Figure 1. Inflation also falls, from just below three percent to a slight rate of deflation. Inflation is relatively smooth because we have no inflation-specific shock in the model.

3.3.1. Heuristics

Before presenting a formal analysis, it may be instructive to see the mechanism that triggers the escape observed in Figure 2. Figure 4 depicts the self-confirming equilibrium in the linearly approximated model. The linearly approximated Taylor-type rule has a slope steeper than the Fisher equation which is a line with slope 1 passing through $-\rho$ on the π-axis. The intersection of the linearized Taylor-type rule and the Fisher equation is the self-confirming equilibrium outcome. From equations (11) and (12), we know that (π_t^d, r_t^d) is distributed around the self-confirming equilibrium. Since the estimated Taylor-type rule (18) must have the same slope as the (true) linearized Taylor-type rule, (π_t^d, r_t^d) must be distributed along the linearized Taylor-type rule.

Since the self-confirming equilibrium is stable, we can find a small neighborhood around the equilibrium in which the gradient of the associated ODE is pointing toward the equilibrium. That is, there exists $\mu > 0$ such that $\forall x \in N_\mu(x^s)$, where x^s is the stationary solution of the associated

FIGURE 4

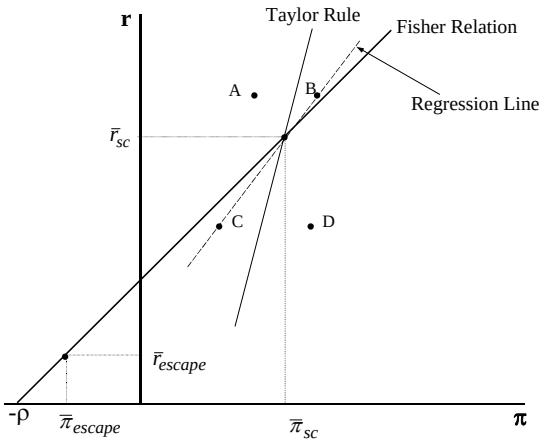

FIG. 4. Heuristic escape dynamics. Data generated along BC leads private sector agents to estimate a less hawkish monetary policy (a regression line with a flatter slope). This causes the inflation target to fall, reinforcing the belief in a less hawkish policy. This process continues until the escape outcome is reached.

ODE, such that

$$\frac{d}{dt}\left|x - x^s\right|^2 < 0. \tag{55}$$

It must be pointed out that the notion of stability does not require that the gradient induced by the ODE points to the self-confirming equilibrium. It suffices that x_t can return to the small neighborhood of the self-confirming equilibrium and remains there after a certain finite period. Indeed, in our case, the path returning to the neighborhood of the self-confirming equilibrium may take a long detour, and (55) generally fails outside of the small neighborhood of the stable solution x^s.

Because the self-confirming equilibrium is stable, we begin by supposing that a cluster of data has been generated about the equilibrium point. We then imagine that for some reason (as we will explain shortly), $\hat{\phi}_{\pi,t}$ has decreased. Despite the fact that the government is actually maintaining the same degree of hawkishness, the *perceived* attitude of the government toward inflation can change. Because the private sector agents conjecture the target inflation rate according to (19), the perceived target inflation

rate also drops. Then, as the government incorporates the private sector's belief through (10), the actual target also shifts toward the left of the self-confirming equilibrium in Figure 4.

Once this lowering of the inflation target has occurred, the new realizations of (π_t^d, r_t^d) will be generated around the new inflation target according to (8) and (9). Since the target inflation and the target interest rate must satisfy the Fisher relation, they must stay along the 45 degree, Fisher relation line passing through $-\rho$ on the π-axis in Figure 4. As the private agents are fitting the regression equation to the observed data, the estimated slope must converge toward 1, because one cluster of data is around the self-confirming equilibrium and the other cluster of data is away from the equilibrium along the Fisher relation, which has slope 1.

As $\hat{\phi}_{\pi,t}$ becomes smaller, $\pi_t^\star$ and $\bar{\pi}_t$ again become still smaller following the same process described above, so that the inflation and nominal interest rate targets are even further away from the self-confirming equilibrium. As most data are accumulated along the Fisher relation, the estimated slope must continue to converge toward 1. The limit of this process has the target inflation rate determined accordingly:

$$\bar{\pi}_t = \pi_t^* = \delta_0 + \delta_1 \tag{56}$$

which is precisely the lower bound of target inflation found in the simulations.

Since $\hat{\phi}_{\pi,t} = 1$ is not a stable point of the associated ODE, the mean dynamics starts to take over to push $\hat{\phi}_{\pi,t}$ back to the self-confirming equilibrium $\hat{\phi}_{\pi,t} = \phi_\pi > 1$.

It remains to explain what kind of a sequence of outcomes can trigger (π_t, r_t) away from the small neighborhood of the self-confirming equilibrium in which the gradient induced by the associated ODE is pointing toward the self-confirming equilibrium. By the definition of the associated ODE, the outcome must stay around the self-confirming equilibrium on average. The weak law of large numbers indicates that the outcome must stay around the mean with a large probability. Essentially, we have to identify a sequence of unusual events that pushes $\hat{\phi}_{\pi,t}$ away from the small neighborhood of

$\phi_\pi > 1$ in order to explain what events can trigger the episode of escape from the self-confirming equilibrium.

Although the simulations were carried out under the assumption that $v_t = (w_t, \eta_t)$ has a Gaussian distribution, it is much more convenient to explain the key intuition of escape if we assume that the perturbations have discrete, binomial distributions. For the sake of discussion, let us assume that η_t can have $\sigma_\eta > 0$ with probability 0.5, and $-\sigma_\eta$ with probability 0.5. Similarly, assume that $\epsilon_t = -\sigma_\epsilon$ or σ_ϵ with an equal probability.

Since w_t and η_t can have two different values respectively, (π_t, r_t) can have four different realizations around the self-confirming equilibrium as depicted in Figure 4 by the points A, B, C, and D. The convex hull of those four realizations forms a parallelogram centered around the self-confirming equilibrium. By connecting the self-confirming equilibrium to each one of the four points, we can see how each realization can change the slope of $\hat{\phi}_{\pi,t}$. Let us call the vector obtained by connecting the self-confirming equilibrium to one of the four realizations of the outcomes a *shifting vector*. We then have four shifting vectors. Since the convex hull of the four points forms a parallelogram, two shifting vectors must be linearly dependent, as they are pointing in opposite directions.

Let us draw a small ball around the self-confirming equilibrium. We need to find the sequence of outcomes that can reach the boundary of the ball with the minimal steps. More precisely, fix a point on the boundary of the ball, and find a sequence of "unusual" events that lead to a small neighborhood of the exit point with minimal steps.

Because the four shifting vectors are two pairs of linearly dependent vectors, if one chooses three or more vectors out of four to generate an escape path from the self-confirming equilibrium, some of the vectors cancel out. In order to minimize the waste of time for escape, any path that can reach a small neighborhood of a fixed point on the boundary of the small ball around the self-confirming equilibrium must be generated by at most two out of four different shifting vectors around the self-confirming equilibrium. A careful examination of Figure 4 reveals that among all six possible combinations of two shifting vectors out of four, exactly one pair of shifting vectors pushes $\hat{\phi}_{\pi,t}$ below ϕ_π. This is the sequence of "unusual"

events that most likely trigger the escape out of the small neighborhood of the self-confirming equilibrium.

One may wonder why the estimated slope $\hat{\phi}_{\pi,t}$ does not escape upward from ϕ_π. To see this, first recall that as the estimated slope becomes larger, the perceived and actual inflation targets also increase. Thus, the center of the distribution of data is shifting toward the *right* of the self-confirming equilibrium in Figure 4. However, as more data are generated along the Fisher relation, which has slope 1, the estimated slope must converge toward 1. This *lowers* the inflation target, tending to move the system back toward the self-confirming equilibrium. Thus, whenever there is an escape, it must happen in such a way that the target inflation *falls* from the self-confirming equilibrium level.

3.3.2. Formal analysis

Because the slope of the estimated Taylor-type rule, $\hat{\phi}_{\pi,t}$ plays a vital role in determining the target inflation rate, we shall focus on the dynamics of $\hat{\phi}_{\pi,t}$. The intercept $\hat{\phi}_{0,t}$ is determined as the regression residual, once $\hat{\phi}_{\pi,t}$ is determined. Let ϕ^s be the self-confirming equilibrium outcome. Similarly, let $\phi^r = (\rho, 1)$. If $\phi = \phi^r$, then the estimated Taylor-type rule coincides with the Fisher relation.

Fix $\rho_\phi > 0$ and define

$$\Omega_{\rho_\phi} = \left\{ \phi_t : \exists T < \infty, \quad \phi_t \notin N_{\rho_\phi}(\phi^s) \quad \& \quad \phi_1 = \phi^s \right\} \tag{57}$$

where $N_{\rho_\phi}(\phi^s)$ is an open ball around ϕ^s with radius ρ_ϕ. By following the analysis of Dupuis and Kushner (1989), one can show that there exists $S^*(\rho_\phi) \in (0, \infty)$ such that

$$-\lim_{a \to 0} a \log \Pr\left(\Omega_{\rho_\phi}\right) \leq S^*(\rho_\phi) \tag{58}$$

which implies that Ω_{ρ_ϕ} is a rare event whose probability vanishes at the rate of $e^{-S^*(\rho_\phi)/a}$, as $a \to 0$, ϕ_t converges to ϕ^s in distribution. Let Σ_v be the covariance matrix for the perturbation v_t.

Proposition 3.1. *$\forall \rho_\phi > 0$,*

$$\lim_{a \to 0} \Pr\left(\phi_t \in N_{\rho_\phi}(\phi^r) \,\middle|\, \Omega_{\rho_\phi}\right) = 1. \tag{59}$$

Proof. As ϕ_t is moving away from ϕ^s, the target $(\overline{\pi}_t, \overline{r}_t)$ also moves along the Fisher relation. Thus, (π_t, r_t) is realized around the target, and the deviation (π_t^d, r_t^d) from the target is bounded by $\overline{v}$ under the first (temporary) assumption. Abusing notation, we can write

$$|r_t - \pi_t - \rho| \leq \overline{v} \qquad \forall t \geq 1. \tag{60}$$

Fix a small $\rho_\phi > 0$. Since the target is moving smoothly with respect to the changes of ϕ_t, we can choose a corresponding $\rho_\pi > 0$ such that $\phi_t \in N_{\rho_\phi}(\phi^s)$ if and only if $\overline{\pi}_t \in N_{\rho_\pi}(\overline{\pi}^s)$. Define $T(\rho_\phi)$ as the first time when

$$\phi_t \notin N_{\rho_\phi}(\phi^s) \tag{61}$$

and similarly, let $T(\rho_\phi/2)$ be the first time when

$$\phi_t \notin N_{\rho_\phi/2}(\phi^s). \tag{62}$$

Since $\phi_{T(\rho_\phi)}$ minimizes the (weighted) forecasting error, $\phi_{T(\rho_\phi)}$ solves

$$\min_{(\phi_0,\phi_\pi)} (1-a) \sum_{j=1}^{T(\rho_\phi)} a^{j-1} \left[r_{T(\rho_\phi)} - \phi_0 - \phi_\pi \pi_{T(\rho_\phi)-j+1} \right]^2. \tag{63}$$

We can write

$$\begin{aligned}
&(1-a) \sum_{j=1}^{T(\rho_\phi)} a^{j-1} \left[r_{T(\rho_\phi)} - \phi_0 - \phi_\pi \pi_{T(\rho_\phi)-j+1} \right]^2 = \\
&\quad (1 - a^{T(\rho_\phi/2)}) \sum_{j=1}^{T(\rho_\phi)-T(\rho_\phi/2)} a^{j-1} \left[r_{T(\rho_\phi)} - \phi_0 - \phi_\pi \pi_{T(\rho_\phi)-j+1} \right]^2 \\
&\quad + a^{T(\rho_\phi)-T(\rho_\phi/2)} \sum_{j=1}^{T(\rho_\phi/2)} a^{j-1} \left[r_{T(\rho_\phi)} - \phi_0 - \phi_\pi \pi_{T(\rho_\phi)-j+1} \right]^2.
\end{aligned} \tag{64}$$

Note that the second term vanishes as $a \to 0$. Instead of the entire objective function, let us focus on the first term of the above equation, and examine

a "simplified" minimization problem:

$$\min_{(\phi_0,\phi_\pi)} (1-a^{T(\rho_\phi/2)}) \sum_{j=1}^{T(\rho_\phi/2)-T(\rho_\phi/2)} a^{j-1}\left[r_{T(\rho_\phi)}-\phi_0-\phi_\pi\pi_{T(\rho)-j+1}\right]^2. \tag{65}$$

Let $\{\hat{\phi}_t^a\}$ be the sequence of estimators obtained from the "original" minimization problem (63) and $\{\hat{\phi}_t^{m,a}\}$ be the sequence obtained from the "simplified" minimization problem (65). We add a to the superscript of ϕ_t in order to emphasize the role of the gain sequence. Since the objective function of (63) converges uniformly to the objective function of (65) which is strictly concave, the sample path $\left\{\hat{\phi}_t^a\right\}$ converges uniformly to $\left\{\hat{\phi}_t^{m,a}\right\}$:

$$\lim_{a\to 0}\left|\hat{\phi}_t^a-\hat{\phi}_a^{m,a}\right|=0 \qquad \forall t\in\{T(\rho_\phi/2),\ldots,T(\rho_\phi)\}. \tag{66}$$

In (65), the agent is fitting a regression line to the data generated around the Fisher relations:

$$E_t\pi_t=\overline{\pi}_t, \quad E_t r_t=\overline{r}_t, \quad \text{and} \quad \overline{r}_t=\overline{\pi}_t+\rho. \tag{67}$$

By the definition of $T(\rho_\phi)$,

$$\left|\overline{\pi}_{T(\rho_\phi)}-\overline{\pi}^s\right|=\rho_\pi \tag{68}$$

which is fixed. Since the feedback rule is a smooth function of ϕ_t, $\overline{\pi}_t$ must be scattered between $\overline{\pi}_{T(\rho_\phi)}$ and $\overline{\pi}^s$, and the number of the data points must increase as $a\to 0$. Hence, as $a\to 0$, the estimated slope $\hat{\phi}_{\pi,T(\rho_\phi)}^{m,a}$ from (65) must converges to the slope of the Fisher relation in distribution:

$$\hat{\phi}_{a,T(\rho_\phi)}^{m,a}\to 1 \tag{69}$$

weakly as $a\to 0$. Combining (66) and (69), we have the desired conclusion. ∎

The proof of Proposition 3.1 reveals the two key elements that generate the escape dynamics. The first element is the fixed gain algorithm that reflects the small amount of suspicion on the part of private sector agents about the stationarity of the underlying economy. With the fixed gain

algorithm, the influence of the past data observed before $T(\rho_\phi/2)$ is depreciated at a geometric rate, which is crucial in obtaining (66). With the least squares estimation in which every data is assigned the equal weight, the data observed before $T(\rho_\phi/2)$ can maintain the same level of influence to $\hat{\phi}^a_{T(\rho_\phi)}$ as $a \to 0$, which keeps the estimator around the self-confirming equilibrium instead of triggering the escape dynamics.

The second element is the misspecification of the model and the feedback. The auxiliary regression (18) presumes ϕ_0 as a constant. However, ϕ_0 reflects the location of the center of the distribution, which is the target inflation and the target nominal interest rate. Since both variables are changing according to the government's feedback rule, the agent's model, especially (18), is misspecified.

However, the misspecified model alone is not enough to trigger the escape dynamics. The unusual dynamics are triggered only when the misspecification is combined with the feedback rule of an agent, in this case, the government. Once the private sector becomes pessimistic about the government's attitude toward inflation (lower $\hat{\phi}_\pi$), the private sector's pessimism is reflected in the government's shift of the target level of inflation, even though its attitude toward inflation ϕ_π remains unchanged. However, as the data is generated along the Fisher relation, the shifted target is interpreted by the private sector as less hawkish attitude of the government, because the estimated slope $\hat{\phi}_\pi$ is lowered toward 1. This process is self-reinforcing until the estimated slope coincides with the slope of the Fisher relation. This is the point where the escape dynamics stops and the mean dynamics takes over to push the private sector's belief back to the self-confirming equilibrium.

4. CONCLUSION

In this paper we have developed a theory of near-zero nominal interest rates, as observed in Japan in the 1990s and the U.S. in the 1930s. Our theory is that the economy inadvertently "slides down a Fisher relation" because of misunderstanding concerning the nature of the government's inflation target. The theory is based on the existence of a self-confirming equilibrium in which inflation and nominal interest rates are relatively high.

Our dynamic system can make sudden departures from that equilibrium towards a persistent low inflation, low nominal interest rate outcome which looks like observed "liquidity trap" episodes in major industrialized countries. These escape dynamics are a consequence of the large deviation properties of our system. We have stressed that three key ingredients are required to generate the escape dynamics. The first of these is that the private sector's model of the government's policy is (subtly) misspecified. The second element is that there is some feedback from beliefs to policy actions. And finally, the private sector needs to learn using a constant gain algorithm, which might be interpreted as allowing these agents to acknowledge their own uncertainty concerning the system in which they operate. With these elements in place, we showed that the long-run behavior of our small macroeconomic model includes recurrent visits to the "liquidity trap" outcome, even though that outcome is not a self-confirming equilibrium of the system.

From the government's point of view, perhaps little can be done to stop the private sector from continually using available data to update their estimates of the policy rule the government uses. And similarly, the nature of the econometric procedure the private sector employs may also be something the government cannot reliably influence. However, the third element needed to generate escape dynamics in this model is the feedback from private sector beliefs to the inflation target. If the government could credibly commit to a constant long-run inflation target, there could be no escape from the unique self-confirming equilibrium in this model. A number of central banks have, in recent years, begun to state their inflation target more explicitly, although not the Bank of Japan or the Federal Reserve.

REFERENCES

1. Benhabib, Jess, Stephanie Schmitt-Grohe, and Martin Uribe. 2001. "The Perils of Taylor Rules." *Journal of Economic Theory*, 96, 40-69.

2. Benveniste, A, M. Metivier and P. Priouret. 1990. *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin.

3. Bullard, James, and Kaushik Mitra. 2002. "Learning About Monetary Policy Rules." *Journal of Monetary Economics*, 49, 1105-1129.

4. Cho, In-Koo, Noah Williams, and Thomas Sargent. 2002. "Escaping Nash Inflation." *Review of Economic Studies*, 69, 1-40.

5. Clarida, Richard, Jordi Gali, and Mark Gertler. 1999. "The Science of Monetary Policy: A New Keynesian Perspective." *Journal of Economic Literature*, XXXVII(4), 1661-1707.

6. Dupuis, Paul, and Harold Kushner. 1989. "Stochastic Approximation and Large Deviations: Upper Bounds and w.p.1 Convergence," *SIAM Journal of Control and Optimization*, 27, 1108-1135.

7. Eusepi, Stefano. 2003. "Forward Versus Backward-Looking Taylor Rules: A 'Global' Analysis." Manuscript, New York University.

8. Evans, George, and Seppo Honkapohja. 2001. *Learning and Expectations in Macroeconomics.* Princeton, New Jersey: Princeton University Press.

9. Evans, George, and Seppo Honkapohja. 2003. "Policy Interaction, Expectations, and the Liquidity Trap." Working Paper, University of Oregon and University of Helsinki.

10. King, Robert, and Charles Plosser, editors. 1999. "Special Issue: Monetary Policy Rules." *Journal of Monetary Economics*, 43, (June).

11. Kushner, Harold. 1984. "Robustness and Approximation of Escape Time and Large Deviation Estimates for Systems with Small Noise Effects." *SIAM Journal of Applied Mathematics*, 44, 160-182.

12. Kushner, Harold, and G.G. Yin. 1997. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag.

13. McCallum, Bennett. 2003. "Multiple solution indeterminacies in monetary policy analysis," *Journal of Monetary Economics*, 50, 1153-1175.

14. Marcet, Albert and Thomas J. Sargent. 1989. "Convergence of Least Squares Learning Mechanisms in Self Referential Linear Stochastic Models." *Journal of Economic Theory*, 48, 337-368.

15. Orphanides, Athanasios, and John Williams. 2003. "The Decline of Activist Stabilization Policy: Natural Rate Misperceptions, Learning, and Expectations." Working paper #2003-24, Federal Reserve Bank of San Francisco.

16. Rotemberg, Julio, and Michael Woodford. 1998. "Interest-Rate Rules in an Estimated Sticky-Price Model." In John Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago Press.

17. Sargent, Thomas. 1999. *The Conquest of American Inflation*. Princeton University Press.

18. Summers, Lawrence. 1991. "How Should Long Term Monetary Policy Be Determined?" *Journal of Money, Credit and Banking*, 23, 625-631.

19. Taylor, John. 1993. "Discretion Versus Policy Rules in Practice." *Carnegie-Rochester Conference Series on Public Policy*, 39, 195-214.

20. Taylor, John. 1999. "A Historical Analysis of Monetary Policy Rules." In John Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago Press.

21. Wheelock, David. 1991. *The Strategy and Consistency of Federal Reserve Monetary Policy 1924-1933*. Cambridge University Press.

22. Wieland, Volker. 2000. "Monetary Policy, Parameter Uncertainty, and Optimal Learning." *Journal of Monetary Economics*, 46, 199-228.

23. Williams, Noah. 2004. "Escape Dynamics in Learning Models." Manuscript, Princeton University.

24. Woodford, Michael. 1990a. "Learning to Believe in Sunspots." *Econometrica*, 58, 277-307.

25. Woodford, Michael. 1990b. "The Optimum Quantity of Money." In B. Friedman and F. Hahn, eds., *Handbook of Monetary Economics*. Volume 2, pp. 1067-1152. Amsterdam: North-Holland.

26. Woodford, Michael. 1999. "Optimal Monetary Policy Inertia." NBER Working Paper #7261, July.

27. Woodford, Michael. 2001. "The Taylor Rule and Optimal Monetary Policy." *American Economic Association Papers and Proceedings*, 91, 232-237.

28. Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press.