Are Low Interest Rates Deflationary? A Paradox of Perfect-Foresight Analysis*

Mariana García-Schmidt
Columbia University

Michael Woodford
Columbia University

September 5, 2015

Abstract

A prolonged period of extremely low nominal interest rates has not resulted in high inflation. This has led to increased interest in the "Neo-Fisherian" proposition according to which low nominal interest rates may themselves cause inflation to be lower. The fact that standard models of the effects of monetary policy have the property that perfect foresight equilibria in which the nominal interest rate remains low forever necessarily involve low inflation (at least eventually) might seem to support such a view. Here, however, we argue that such a conclusion depends on a misunderstanding of the circumstances under which it makes sense to predict the effects of a monetary policy commitment by calculating the perfect foresight equilibrium consistent with the policy. We propose an explicit cognitive process by which agents may form their expectations of future endogenous variables. Under some circumstances, such as a commitment to follow a Taylor rule, a perfect foresight equilibrium (PFE) can arise as a limiting case of our more general concept of reflective equilibrium, when the process of reflection is pursued sufficiently far. But we show that an announced intention to fix the nominal interest rate for a long enough period of time creates a situation in which reflective equilibrium need not resemble any PFE. In our view, this makes PFE predictions not plausible outcomes in the case of policies of the latter sort. According to the alternative approach that we recommend, a commitment to maintain a low nominal interest rate for longer should always be expansionary and inflationary, rather than causing deflation; but the effects of such "forward guidance" are likely, in the case of a long-horizon commitment, to be much less expansionary or inflationary than the usual PFE analysis would imply.

*We would like to thank Gauti Eggertsson, Jamie McAndrews, Rosemarie Nagel, Jon Steinsson and Lars Svensson for helpful comments, and the Institute for New Economic Thinking for research support.

1 Perfect-Foresight Analyses of the Effects of Forward Guidance: A Paradox

One of the more notable features of recent monetary experience has been the fact that first Japan, and now more recently the U.S. as well, have gone through prolonged periods of extremely low nominal interest rates (overnight interest rates reduced practically to zero and kept there for years) without this leading to the sort of inflationary spiral that one might have expected to follow from such a reckless experiment. Instead, inflation has remained low, below both countries' desired levels of inflation (and even below zero, much of the time, in Japan), while real activity has remained disappointing as well. A common reaction to these surprising developments has been to conclude that financial crises of the kind that both countries experienced can lower the equilibrium real rate of interest for a very prolonged period of time, so that real interest rates that seem very low by historical standards may nonetheless continue to be contractionary.

But some have proposed an alternative interpretation of these experiences, according to which low nominal interest rates themselves may *cause* inflation to be lower. In this view, the monetary policy reactions to these crises may have actually prolonged the disinflationary slumps by creating disinflationary expectations. Under such a view, actually promising to keep interest rates low for a longer period than would otherwise have been expected — as both the Fed and a number of other central banks have done in the recent period[1] — would be the worst possible policy for a central bank worried that inflation will continue to run below its target, and some (beginning with Bullard, 2010, and Schmitt-Grohé and Uribe, 2010) have proposed that such a central bank should actually *raise* interest rates in order to head off the possibility of a deflationary trap. As the period over which the U.S. has kept its federal funds rate target near zero has continued, views of this kind, that some have taken to calling "neo-Fisherian," have gained increasing currency, at least on the internet.[2]

Moreover, it might seem that even a standard textbook model of the effects of alternative monetary policy commitments would support the "neo-Fisherian" position. The most straightforward theoretical argument proceeds in two steps.[3] One

[1] See, for example, Woodford (2012) for a discussion of these experiences.

[2] See, for example, Cochrane (2015b) for discussion and additional references.

[3] The argument is explained more formally in section 2.2 below.

first considers what should happen if a central bank were to commit to maintain the short-term nominal interest rate at an arbitrarily chosen level *forever*. According to a traditional view, famously articulated by Friedman (1968), this is not a possible experiment, because any such attempt would lead to explosive inflation dynamics that would require the central bank to abandon the policy in finite time. But in fact, many modern equilibrium models of inflation determination, including standard New Keynesian models, imply that there exist rational-expectations equilibria associated with such a policy in which inflation and other variables remain forever bounded — so that there is no reason to deny the logical possibility of the proposed thought experiment.[4] In a deterministic setting, there is typically a one-dimensional continuum of perfect foresight equilibria consistent with this policy commitment, all of which converge asymptotically to a steady state in which the constant inflation rate is the one determined by the nominal interest-rate target and the Fisher equation. Thus one might conclude that such an experiment should lead to an inflation rate that converges to the one determined by the Fisher equation (and hence that is higher by one percentage point for each percentage point increase in the nominal interest-rate target), at least eventually.

The second step in the argument notes that it doesn't make sense to suppose that the outcome resulting from a given forward path for policy should be extremely sensitive to small changes in anticipated policy that relate only to the very distant future. More specifically, one might assert that an expected shift in the monetary policy rule should have an effect on outcomes now that shrinks to zero as the date of the anticipated policy shift is pushed far enough into the future.[5] But this means that a commitment to keep the nominal interest rate at some level up until some finite date T should not have consequences that are very different than those that would follow from keeping the interest rate at that level forever. If keeping the interest rate low forever must eventually *lower* the inflation rate, then there must be some finite length of time such that keeping the interest rate low for that length of time *also* must eventually lower the inflation rate almost as much. It is only a question of how long a period of low interest rates should be required to observe this effect.

This is a paradoxical result: it seems that the very assumptions that underly com-

[4]This is emphasized in expositions of the neo-Fisherian view such as that of Cochrane (2015b).

[5]This is the basis for the proposal in Cochrane (2015a) that a plausible analysis should select the "backward stable" perfect foresight solution consistent with a given forward path policy.

mon arguments for the efficacy of forward guidance — the use of a New Keynesian model of the monetary transmission mechanism, and the assumption of perfect foresight (or rational expectations) to determine the effects of a given policy commitment — imply that a commitment to keep interest rates low for a long time should be even more disinflationary than a plan of returning sooner to a more normal policy. Yet this is not at all what standard model-based analyses of the implications of forward guidance have concluded, and it is certainly not what policymakers have assumed when recently announcing or contemplating commitments of that kind.

It might seem that an argument of the kind just sketched about the consequences of policies expected to last for unboundedly long periods of time has no consequences for anything we will ever actually observe, and therefore no bearing upon either practical policy analysis or the interpretation of historical experience. But the standard approach to analyzing the consequences of an expectation that the short-term interest rate will remain at the zero lower bound (ZLB) for several more quarters — which looks at the perfect foresight equilibrium (or the rational-expectations equilibrium, in the case of a stochastic model) consistent with the forward path of policy that converges asymptotically to the steady state in which the central bank's long-run inflation target is achieved — has the consequence, in a standard (very forward-looking) New Keynesian model, that as the length of time that the interest rate is expected to remain at zero is made longer, the predicted positive effects on inflation and output at the time that the policy attention is announced grow explosively, as shown by Del Negro *et al.* (2013), Chung (2015), and McKay *et al.* (2015). This prediction violates the principle that anticipated policy paths that differ only in the specification of policy far in the future should have similar near-term effects; but if one thinks that the conclusion must be wrong about the effects of commitments to long spells of zero-interest-rate policy, one may suspect that it is wrong about the effects of shorter-range policy commitments as well.

And similarly, if one thinks that selecting instead the "backward stable" perfect foresight equilibrium as the relevant model prediction (as proposed by Cochrane, 2015a) makes sense in the case of commitments to a long spell of zero-interest-rate policy, one may find this a reason to regard it as the more sensible prediction in the case of shorter-range policy commitments as well. But the conventional equilibrium selection and the "backward stable" selection lead to very different predictions about

the effects of even periods of modest length at the zero lower bound.[6]

Thus the conclusion that one reaches about the paradoxes resulting from attempts to analyze very long spells at the zero lower bound matters for the analysis that one should give of types of policy experiments that have recently been attempted or contemplated. Indeed, if one accepts the analysis proposed by Cochrane (2015a), the neo-Fisherian logic applies also to spells at the zero lower bound of only a few years. In the numerical solutions that he displays, a temporary reduction of the natural rate of interest to a level that makes the zero lower bound inconsistent with the central bank's inflation target — so that the ZLB requires an interest rate higher than the one consistent with the target inflation rate, for a time — is *inflationary*, rather than deflationary as in analyses like that of Eggertsson and Woodford (2003). And maintaining a higher interest rate during the period of the shock would be *even more* inflationary, according to the "backward stable" equilibrium selection.

In this paper we consider whether a standard New Keynesian model of the effects of monetary policy requires one to accept paradoxical conclusions of this kind.[7] We shall argue that it does not. Our quarrel, however, is not with the postulate that anticipated changes in policy sufficiently far in the future should have negligible effects on current economic outcomes. Rather, we deny the practical relevance of the perfect foresight solutions (or more generally, rational-expectations solutions) of the model under the thought experiment of a permanent interest-rate peg.

Moreover, our criticism of the perfect-foresight analysis of this case is not based on a wholesale denial of the plausibility of forward-looking expectations. It is well-known that Friedman's view of the consequences of an interest-rate peg can be defended if one supposes that people's expectations are purely backward-looking, as Friedman's

[6] See the demonstration of this in Cochrane (2015), sec. 3.1

[7] Of course, we do not pretend to consider all of the logically possible models, and all of the logically possible assumptions about policy, that might be consistent with neo-Fisherian claims. For example, we do not discuss Cochrane's (2014) derivation of neo-Fisherian conclusions under the assumption of a non-Ricardian fiscal policy; here we are solely concerned with situations in which fiscal policy is expected to be Ricardian, in a sense made precise in Woodford (2013). We would dispute the argument in Cochrane (2014) that a non-Ricardian fiscal policy should be assumed because the path of the price level is otherwise indeterminate in New Keynesian models. We offer here a way of obtaining a determinate prediction despite Ricardian expectations regarding fiscal policy, and show that it leads to quite different conclusions from those that would result from the kinds of expectations about fiscal policy analyzed in Cochrane (2014), in addition to differing from the predictions obtained in Cochrane (2015a) by selecting the "backward stable" equilibrium.

informal discussion presumed.[8] However, this particular defense of conventional views about the effects of interest-rate policy would *also* imply that "forward guidance" as to a central bank's intentions regarding future policy should have no effects on equilibrium outcomes, as expectations regarding the future are assumed to follow solely from the data that have already been observed. Such a view would imply that if the zero lower bound prevents a central bank from lowering the current short-term rate enough to achieve its stabilization objectives through that channel alone, there is nothing further to be done; keeping the interest rate low even beyond the end of the period in which the bank's current targets cannot be achieved can achieve higher output and inflation in that later period, but because this will not be anticipated until it occurs, this will do nothing to improve outcomes during the constrained period (while meaning less successful stabilization later). So while it would not imply that a commitment to keep the nominal interest rate low for a long time would actually lower inflation, it would nonetheless imply that such a policy would impede macroeconomic stabilization rather than improving it.

We offer a different reason for rejecting the neo-Fisherian conclusion. We believe that people are at least somewhat forward-looking; this is why central bank commitments about the way in which monetary policy will be conducted in the future (such as explicit inflation targets) matter. Nonetheless, it may not be reasonable to expect that the outcome associated with a given policy commitment should be a perfect foresight equilibrium, even when the commitment is fully credible and people have the knowledge about how the economy works that would be required for calculation of such an equilibrium.

We argue that predicting what should happen as a result of a particular policy commitment requires that one model the cognitive process by which one imagines people to arrive at particular expectations taking that information into account. In this paper, we offer a simple example of such an explicit model of reasoning. Under our approach, a perfect foresight equilibrium (or more generally, a rational-expectations equilibrium[9]) can be understood as a limiting case of a more general concept of re-

[8] One can show formally, in a model derived from intertemporal optimization of the kind used below, that an interest-rate peg will imply explosive dynamics if expectations are based on extrapolation from past data, as under Friedman's hypothesis of "adaptive expectations" or the hypothesis of "least-squares learning" that has been popular more recently. See Woodford (2003, sec. 2.3) for discussion and references.

[9] We consider only deterministic environments in which, after some (possibly unexpected) change

flective equilibrium, which limit may be reached under some circumstances if the process of reflection about what forward paths for the economy to expect is carried far enough. Our concept of reflective equilibrium is similar to the "calculation equilibrium" proposed by Evans and Ramey (1992, 1995, 1998): we consider what economic outcomes should be if people optimize on the basis of expectations that they derive from a process of reflection about what they should expect, given both their understanding of how the economy works and (as part of that structural knowledge) their understanding of the central bank's policy intentions.

Furthermore, like Evans and Ramey, we model this process of reflection as an iterative process that adjusts the provisional forecasts that are entertained at a given stage of the process in response to the predictable discrepancy between those forecasts and what one should expect to happen if people were to behave optimally on the basis of those forecasts. Thus the process is one under which beliefs should continue to be adjusted, if the process is carried farther, unless perfect-foresight equilibrium beliefs have been reached. And like Evans and Ramey, we are interested in the theoretical question of where such a process of belief revision would end up asymptotically, if carried forward indefinitely, but we regard it as more realistic to suppose that in practice, the process of reflection will be suspended after some finite degree of reflection, and people will act upon the beliefs obtained in this way.

The most important difference between our approach and that of Evans and Ramey is that the primary goal of their analysis is to determine how far the belief revision process should be carried forward, by specifying costs of additional calculation and a criterion for judging the benefits that should be weighed against those costs; we do not propose any explicit model of such costs or the decision to terminate the process of belief revision. Our concerns are instead to determine whether the process will necessarily reach a perfect foresight equilibrium even if carried forward indefinitely; to ask which perfect foresight equilibrium is reached in the case that

in economic fundamentals and/or the announced path of monetary policy, neither fundamentals nor policy should depend on any further random events, and so we consider only the reasonableness of assuming a perfect foresight equilibrium. But the kind of reflective equilibrium that we define below could also be considered in stochastic environments, in which case we could instead consider under what conditions the process of reflection will eventually converge to a rational-expectations equilibrium; and some of the convergence results obtained below have direct extensions to stochastic environments. To economize on notation and technicality, however, we here expound the idea only in the simpler deterministic setting.

the process converges; and to understand what determines the speed of convergence when it occurs.

In our view, the predictions obtained by considering the perfect foresight equilibrium (PFE) consistent with a given forward path for policy are of practical relevance only in that the belief revision process converges to those PFE beliefs, sufficiently rapidly and from a large enough range of possible beliefs that may be initially entertained. If one has fast convergence to certain PFE beliefs from many possible starting points, then the PFE predictions should be a good approximation to what one should expect to occur in a reflective equilibrium, under any of a considerable range of assumptions about where the process starts and how far it is carried forward. We show below that standard conclusions about equilibrium determination under a Taylor rule (when the zero lower bound does not constrain policy) can be justified in this way; our analysis not only provides a reason for interest in the perfect-foresight (or rational-expectations) predictions about such a policy commitment, but explains why one particular PFE solution should be regarded as the relevant prediction of the model, addressing Cochrane's (2011) critique of the standard New Keynesian literature.

If, instead, a particular perfect foresight equilibrium cannot be reached under the belief revision process, except by starting from extremely special initial beliefs, then we do not think it is plausible to expect actual outcomes to resemble those PFE outcomes.[10] And if the belief revision process does not converge, or if it converges only very slowly, then we do not believe there is ground to make any very specific prediction about the beliefs that people should be expected to hold in practice and hence about the economic outcomes that should be observed; and the range of outcomes that should be considered to represent reasonable possibilities need have little to do with the set of perfect foresight equilibria. We show below that in a standard New Keynesian model, the thought experiment of an interest-rate peg that is maintained forever produces a situation of this kind: while perfect foresight equilibria do indeed exist, the belief revision process that we consider does not converge to any of them, and the set of reflective equilibria resulting from different finite degrees of reflection do not resemble perfect foresight equilibria. Thus in our view, the forward guidance paradox sketched above results from reliance upon the concept of perfect foresight

[10]This is our view of the "backward stable" PFE solutions analyzed by Cochrane (2015a) in the case of a temporary interest-rate peg.

equilibrium in a context in which it is especially inappropriate.

Some may protest that an equilibrium concept that allows no definite conclusion about what should occur (in the case of non-convergence or slow convergence) is of no use in informing policy design. Yet as our results below illustrate, even in such a case, it may well be possible to derive *qualitative* conclusions about the effects on a reflective equilibrium that should be expected from changing policy in a particular direction; and these may differ, even as to sign, from those that would be suggested by considering the set of perfect foresight equilibria.

In particular, we show that in our model, a commitment to maintain a low nominal interest rate for a longer period of time — or to maintain a lower rate, for any fixed length of time — will typically result (under any given finite degree of reflection) in increased aggregate demand, increasing both output and inflation in the near term, though the exact degree of stimulus that should result depends (considerably) on the assumed degree of reflection. This is true regardless of the length of time for which the interest-rate peg is expected to be maintained, and even in the limit of a perpetual interest-rate peg. Thus consideration of the reflective equilibrium resulting from a finite degree of reflection yields conventional conclusions about the sign of the effects of commitments to lower interest rates in the future, and does so without implying any non-negligible effects of changing the specification of policy only very far in the future.

Hence the reflective equilibrium analysis avoids both of the paradoxical conclusions that a PFE analysis requires one to choose between: affirming either that maintaining low nominal interest rates must eventually be deflationary, or that the outcome implied by a given policy commitment can depend critically on the specification of policy extremely far in the future. It implies that PFE (or rational-expectations equilibrium) analyses of the effects of committing to keep the nominal interest rate low for a longer (but still fairly short) period of time, under the conventional approach to equilibrium selection, are likely to be correct as to the *signs* of the predicted effects, but that the numerical magnitudes of the effects obtained from such analyses may be quite inaccurate. In particular, our numerical illustrations below suggest that the predicted effects on output and inflation from the PFE analysis are likely to be upper bounds on the effects that should occur in a reflective equilibrium with only a finite degree of reflection — and indeed, wild exaggerations in cases where the interest rate is expected to remain at the zero lower bound for many years.

We proceed as follows. Section 2 introduces our New Keynesian model of inflation and output determination under alternative monetary policies and alternative assumptions about private-sector expectations, and the belief revision process that underlies our proposed concept of reflective equilibrium. Section 3 then considers reflective equilibrium when the forward path of monetary policy is specified by a Taylor rule, and both the path of policy and the economy's exogenous fundamentals are such that the ZLB never binds in equilibrium. We allow the Taylor rule to involve a possibly time-varying intercept (or inflation target), so that we can analyze a type of forward guidance (but not one that involves commitment to a constant interest rate). By contrast, section 4 considers reflective equilibrium in the less well-behaved case of an expectation that the short-term interest rate will remain fixed until some horizon T, and then revert to a Taylor rule thereafter; and also the limiting case of a commitment to keep the interest rate fixed forever. Section 5 offers concluding reflections.

2 Reflective Equilibrium in a New Keynesian Model

We expound our concept of reflective equilibrium in the context of a log-linearized New Keynesian (NK) model. The model is one that has frequently been used, under the assumption of perfect foresight or rational expectations, in analyses of the potential effects of forward guidance when policy is temporarily constrained by the zero lower bound (e.g., Eggertsson and Woodford, 2003; Werning, 2012; McKay *et al.*, 2015; Cochrane, 2015a).[11] As in the analyses of Evans and Ramey (1992, 1995, 1998), we must begin by specifying the *temporary equilibrium relations* that map arbitrary subjective expectations about future economic conditions into market outcomes; these relations play a crucial role in the process of reflection that we wish to model, in addition to being required in order to predict what should happen if people's beliefs do not converge to PFE beliefs. Because these relations are not generally discussed in a form that would be valid under arbitrary subjective expectations in expositions of the NK model that consider only its rational-expectations equilibria, it is necessary to briefly sketch the foundations of the model. The presentation here

[11] Werning (2012) and Cochrane (2015a) analyze a continuous-time version of the model, but the structure of the model that they consider is otherwise the same as the discrete-time model considered here.

largely follows Woodford (2013), where the derivations are discussed in more detail.

2.1 Temporary Equilibrium Relations

In our model, both households and firms solve infinite-horizon decision problems, and hence their optimal decision rules depend on their expectations about economic conditions in a series of future periods extending indefinitely. It is important that we explicitly represent the way in which actions depend on expectations about different future dates, because of our interest in analyzing the effects of announcements about future policy that may refer to points in time at different distances from the present.

The economy is made up of identical, infinite-lived households. Each household i seeks to maximize a discounted flow of utility

$$\hat{E}_t^i \sum_{T=t}^{\infty} e^{\sum_{s=t}^{T-1} \rho_s} [u(C_T^i) - v(H_T^i)] \tag{2.1}$$

when planning their path of consumption, looking forward from date t. Here C_t^i is a Dixit-Stiglitz aggregate of the household's purchases of differentiated consumer goods, H_t^i is hours worked, the sub-utility functions satisfy $u' > 0, u'' < 0, v' > 0, v'' \geq 0$, and ρ_t is a possibly time-varying discount rate. We allow for the possibility of a non-uniform discount rate in order to introduce a reason why the ZLB may temporarily constrain monetary policy; the fact that intra-temporal preferences are uniform over time will allow the efficient level of output to be constant over time.[12] The operator $\hat{E}_t^i$ indicates that this objective is evaluated using the future paths of the variables implied by the household's subjective expectations, which need neither be model-consistent nor common across all households.

For simplicity, there is assumed to be a single traded asset each period: one-period riskless nominal debt (a market for which must exist in order for the central bank to control a short-term nominal interest rate). Each household also owns an equal share of each of the firms (discussed below), but these shares are assumed not to be tradeable. In the present exposition, we abstract from fiscal policy, by assuming that there are no government purchases, government debt, or taxes and transfers.[13] We can then define the set of expenditure sequences $\{C_T\}$ for dates $T \geq t$ that the

[12] In Woodford (2013), a more general version of the model is presented, in which a variety of other types of exogenous disturbances are allowed for.

[13] Woodford (2013) shows how the temporary equilibrium framework can be extended to include

household expects will be feasible as a function of the expected paths of real income, the one-period nominal interest rate, and the rate of inflation.

We can then solve for the household's optimal expenditure plan as a function of those expectations, and log-linearize the optimal decision rule around the constant plan that is optimal in the event that $\rho_T = \bar{\rho} > 0$ for all $T \geq t$, the inflation rate is expected to equal the central bank's target rate π^* in all periods, and real income and the nominal interest rate are also expected to be constant in all periods at values that represent a PFE for a monetary policy that achieves the inflation target at all times.[14] We obtain

$$c_t^i \;=\; \sum_{T=t}^{\infty} \beta^{T-t}\, \hat{E}_t^i \left\{(1-\beta) y_T \;-\; \beta\sigma \left(i_T - \pi_{T+1} - \rho_T\right)\right\} \tag{2.2}$$

where $\{y_T, i_T, \pi_T\}$ are the expected paths of real income (or aggregate output, in units of the Dixit-Stiglitz aggregate), the nominal interest rate, and inflation, and all variables appearing in the equation are measured as log deviations from their steady-state values (hence the use of c_t^i rather than the C_t^i that appears in (2.1)). Here $\beta \equiv e^{-\bar{\rho}} < 1$ is the steady-state discount factor and $\sigma > 0$ is the intertemporal elasticity of substitution of consumer expenditure. Note that (2.2) generalizes the familiar "permanent-income hypothesis" formula (obtained by keeping only the y_T terms on the right-hand side) to allow for a non-constant desired path of spending owing either to variation in the anticipated real rate of return or transitory variation in the rate of time preference.

We assume that households can correctly forecast the variation over time (if any) in their discount rate in their intertemporal planning, so that $\hat{E}_t^i \rho_T = \rho_T$ for all $T \geq t$.[15] The subjective expectations (that instead may not be model-consistent)

fiscal variables. The resulting temporary equilibrium relations are essentially of the kind derived here, as long as households have "Ricardian expectations" (defined precisely there) regarding their future net tax liabilities: that is, they expect that no matter how prices and interest rates evolve, net taxes collected by the government will have a present value exactly equal to the value of outstanding government debt. The assumption of Ricardian expectations is important for one's conclusions about the macroeconomic effects of interest-rate policy, as shown by Cochrane (2014).

[14] We assume that $\pi^* > -\bar{\rho}$, so that the required nominal interest rate in this steady state is positive.

[15] This means that expectations regarding future preference shocks are treated differently in (2.3) below than in the expression given in Woodford (2013). The definition of the composite expectational variable v_t^i is correspondingly different.

regarding future conditions that matter for a household's expenditure decision can then be collected in a single expectational term, allowing us to rewrite (2.2) as

$$c_t^i \;=\; (1-\beta)y_t \;-\; \beta\sigma i_t \;+\; \beta g_t \;+\; \beta\,\hat{E}_t^i v_{t+1}^i, \tag{2.3}$$

where

$$g_t \;\equiv\; \sigma \sum_{T=t}^{\infty} \beta^{T-t}\,\rho_T$$

measures the cumulative impact on the urgency of current expenditure of a changed path for the discount rate, and

$$v_t^i \;\equiv\; \sum_{T=t}^{\infty} \beta^{T-t}\,\hat{E}_t^i\,\{(1-\beta)y_T \;-\; \sigma(\beta i_T - \pi_T)\}$$

is a household-specific subjective variable.

Then defining aggregate demand y_t (which will also be aggregate output and each household's non-financial income) as the integral of expenditure c_t^i over households i, the individual decision rules (2.3) aggregate to an *aggregate demand (AD) relation*

$$y_t \;=\; g_t \;-\; \sigma i_t \;+\; e_{1t}, \tag{2.4}$$

where

$$e_{1t} \;\equiv\; \int \hat{E}_t^i v_{t+1}^i\,di$$

is a measure of average subjective expectations.

The continuum of differentiated goods are produced by Dixit-Stiglitz monopolistic competitors, who each adjust their prices only intermittently to changing market conditions; as in the Calvo-Yun model of staggered pricing, only a fraction $1-\alpha$ of prices are reconsidered each period, where $0 < \alpha < 1$ measures the degree of price stickiness. Our version of this model differs from many textbook presentations (but follows the original presentation of Yun, 1996) in assuming that prices that are not reconsidered in any given period are automatically increased at the target rate π^*.[16] If a firm j reconsiders its price in period t (rather than simply increasing it at the

[16]This allows us to assume a positive steady-state inflation rate — which is important for the quantitative realism of the numerical examples below, since the steady-state inflation rate matters for the tightness of the ZLB constraint — while at the same time retaining the convenience of a steady state in which the prices of all goods are identical, despite the assumption of staggered pricing.

rate π^*), it chooses a price that it expects to maximize the present discounted value of profits in all future states prior to the next reconsideration of its price, given its subjective expectations regarding the evolution of aggregate demand $\{y_T\}$ for the composite good and of the log Dixit-Stiglitz price index $\{p_T\}$ for all $T \geq t$. A log-linear approximation to its optimal decision rule (again around the steady state with constant inflation rate π^*) takes the form

$$p_t^{*j} = (1-\alpha\beta)\sum_{T=t}^{\infty}(\alpha\beta)^{T-t}\hat{E}_t^j[p_T + \xi y_T - \pi^*(T-t)] \tag{2.5}$$

$$- (p_{t-1} + \pi^*) \tag{2.6}$$

where p_t^{*j} is the amount by which $j's$ log price exceeds the average of the prices that are not reconsidered, $p_{t-1} + \pi^*$, $\xi > 0$ measures the elasticity of a firm's optimal relative price with respect to aggregate demand,[17] and the operator $\hat{E}^j[\cdot]$ indicates that what matter are the subjective expectations of firm j regarding future market conditions.

Again, the terms on the right-hand side of (2.6) involving subjective expectations of conditions at various future horizons can be collected in a single composite term, $\alpha\beta\hat{E}_t^j p_{t+1}^{*j}$. Aggregating across the prices chosen in period t by the continuum of firms, we obtain an implied *aggregate supply (AS) relation*

$$\pi_t = \kappa y_t + (1-\alpha)\beta\, e_{2t} \tag{2.7}$$

where $\pi_t \equiv p_t - p_{t-1} - \pi^*$ is inflation in excess of the target rate,

$$\kappa \equiv \frac{(1-\alpha)(1-\alpha\beta)\xi}{\alpha} > 0,$$

and

$$e_{2t} \equiv \int \hat{E}_t^j p_{t+1}^{*j}\, dj$$

measures average expectations of the composite variable.

We can close the system by assuming a reaction function for the central bank of the Taylor (1993) form

$$i_t = \bar{\imath}_t + \phi_\pi \pi_t + \phi_y y_t \tag{2.8}$$

[17]The parameter ξ is thus a measure of the degree of "real rigidities." See Woodford (2003, chap. 3) for a detailed discussion of its dependence on underlying parameters relating to preferences, technology and market structure.

where the response coefficients satisfy $\phi_\pi, \phi_y \geq 0$. We allow for a possibly time-varying intercept in order to consider the effects of announcing a transitory departure from the central bank's normal reaction function. (More generally, we consider below the possibility of situations in which the response coefficients are assumed to be different at different times, though they are assumed time-invariant in (2.8).) Equations (2.4), (2.7) and (2.8) then comprise a three-equation system, that determines the temporary equilibrium (TE) values of y_t, π_t, and i_t in a given period, as functions of the exogenous disturbances $(g_t, \bar{\imath}_t)$ and subjective expectations (e_{1t}, e_{2t}). Under our sign assumptions, it is easily shown that the TE values are uniquely determined, linear functions of the vector of disturbances and the vector of subjective expectations (see the Appendix for details).

As preparation for our discussion of the process of expectation revision, it is useful to note the relationship between subjective expectations and the actual values of the variables that people seek to forecast. The two sufficient statistics for subjective expectations e_{it} are each the average forecast of a certain average of the future values of a certain composite variable a_{iT} at different horizons $T > t$, where a_{iT} is a variable that is determined purely by disturbances and subjective expectations in period T. For $i = 1, 2$, we can write

$$e_{it} = (1 - \delta_i) \sum_{j=0}^{\infty} \delta_i^j \bar{E}_t a_{i,t+j+1}, \tag{2.9}$$

where

$$\delta_1 = \beta, \qquad \delta_2 = \alpha\beta$$

(so that $0 < \delta_i < 1$ for both variables),

$$\begin{aligned} a_{1t} &\equiv y_t - \frac{\sigma}{1-\beta}(\beta i_t - \pi_t), \\ a_{2t} &\equiv \frac{1}{1-\alpha\beta}\pi_t + \xi y_t, \end{aligned}$$

and the operator $\bar{E}_t[\cdot]$ indicates the average of the population's forecasts at date t.[18]

[18] While we still allow for the possibility of heterogeneous forecasts, from here on we assume that the distribution of forecasts across the continuum of households is the same as the distribution across the continuum of firms, and so refer simply to the distribution of forecasts. $\bar{E}_t[\cdot]$ refers to the mean of this distribution of forecasts at some date t.

We can then use the TE relations to solve for the equilibrium values of the variables a_{it} that people seek to forecast as linear functions of the current vector of disturbances and current average expectations. This solution can be written in the form

$$a_t \;=\; M\, e_t \;+\; m\, \omega_t, \tag{2.10}$$

where a_t is the vector consisting of (a_{1t}, a_{2t}), e_t is the vector consisting of (e_{1t}, e_{2t}), ω_t is the vector consisting of $(g_t, \bar{\imath}_t)$, and the matrices of coefficients are given in the Appendix. The system (2.10) shows how expectations determine the endogenous variables that are themselves being forecasted in those expectations, as indicated by (2.9).

2.2 Perfect Foresight Equilibrium

The assumption of perfect foresight equilibrium adds to the above model the further assumption that the expected paths for output, inflation and the interest rate (and hence the expected paths for the variables $\{a_t\}$) are precisely the paths for those variables implied by the TE relations under those expectations. Thus a PFE corresponds to sequences $\{a_t, e_t\}$ that both satisfy (2.10) each period and satisfy (2.9) when the equilibrium paths $\{a_t\}$ are substituted for the average expectations in those equations.

It can be shown (see Woodford, 2013) that under the PFE assumption, the TE relations (2.4) and (2.7) imply that the paths of output, inflation and the interest rate must satisfy difference equations of the form

$$y_t = y_{t+1} - \sigma(i_t - \pi_{t+1} - \rho_t) \tag{2.11}$$

$$\pi_t = \kappa y_t + \beta \pi_{t+1} \tag{2.12}$$

which are simply perfect-foresight versions of the usual "New Keynesian IS curve" and "New Keynesian Phillips curve" respectively. Using the policy specification (2.8) to eliminate i_t, one obtains a pair of difference equations that can be written in the form

$$x_t \;=\; B\, x_{t+1} + b\,(\rho_t - \bar{\imath}_t) \tag{2.13}$$

where x_t is the vector consisting of (y_t, π_t), and the matrix B and vector b are defined in the Appendix.

Under our sign assumptions for the model coefficients, the matrix B is invertible, and the system (2.13) can be uniquely solved for x_{t+1} as a function of x_t and the period t disturbances. One then obtains a two-parameter family of possible PFE solutions consistent with any given forward paths for the disturbances, corresponding to the set of possible choices for the elements of x_0. The asymptotic behavior of these solutions as t is made large depends as usual on the eigenvalues of the matrix B.

As shown in the Appendix, the matrix B has both eigenvalues inside the unit circle if and only if

$$\phi_\pi + \frac{1-\beta}{\kappa}\phi_y \;>\; 1 \tag{2.14}$$

so that the "Taylor Principle" is satisfied. In this case, there is a unique bounded PFE solution for the sequences $\{x_t\}$ corresponding to any bounded sequences $\{\rho_t, \bar{\imath}_t\}$, obtained by "solving forward" the system (2.13) to obtain

$$x_t \;=\; \sum_{j=0}^{\infty} B^j b\,(\rho_{t+j} - \bar{\imath}_{t+j}). \tag{2.15}$$

When this is uniquely defined, we shall call this the "forward stable" PFE (FS-PFE). It is common to regard this as the relevant prediction of the model in such a case;[19] below we shall provide a justification for this in terms of our concept of reflective equilibrium.

This solution implies that in the case of a sufficiently transitory change in policy, a reduction of $\bar{\imath}_t$ (for a given path of the real disturbance) must be both expansionary and inflationary (must raise both y_t and π_t), while the nominal interest rate is temporarily reduced (though by less than the reduction in $\bar{\imath}_t$). In the case of a sufficiently persistent shift in $\bar{\imath}_t$, output, inflation and the nominal interest rate are *all* predicted to increase, because of the endogenous effect of the output and inflation increases on the central bank's interest-rate target;[20] but even in this case, a downward shift in the reaction function (reducing the interest-rate target implied by any given current levels of inflation and output) is inflationary rather than deflationary.

[19] See however Cochrane (2011) for objections to this interpretation.

[20] In a more realistic model than the simple NK model used in this paper, there will be delays in the effect of the policy change on output and inflation. It is then possible to have an initial decline in the nominal interest rate in the case of an expansionary monetary policy shock, even in the case of a relatively persistent shift in the central-bank reaction function, as shown in Woodford (2003, secs. 5.1-5.2).

If the inequality in (2.14) is reversed, the matrix B instead has two real eigenvalues satisfying

$$0 < \mu_1 < 1 < \mu_2,$$

so that the larger is outside the unit circle. In particular, this is true if the central bank fixes the forward path for the nominal interest rate (the case $\phi_\pi = \phi_y = 0$), regardless of whether this path is constant. In this case, there is no longer a unique bounded solution; instead, assuming again that the sequences $\{\rho_t, \bar{\imath}_t\}$ are bounded, there is a bounded PFE solution

$$x_t = v_1(e_1'b) \sum_{j=0}^{\infty} \mu_1^j \left(\rho_{t+j} - \bar{\imath}_{t+j}\right) \; - \; v_2(e_2'b) \sum_{j=1}^{t} \mu_2^{-j} \left(\rho_{t-j} - \bar{\imath}_{t-j}\right) \; + \; \chi \, v_2 \mu_2^{-t} \quad (2.16)$$

in the case of any real number χ. (In this expression, v_i is the right eigenvector corresponding to eigenvalue μ_i, e_i' is the left eigenvector corresponding to that same eigenvalue, and we normalize the eigenvectors so that $e_i'v_i = 1$ for each i.) Since such solutions necessarily exist, the PFE analysis gives us no reason to suppose that there is anything problematic about a commitment to fix a path for the nominal interest rate, including a commitment to fix it at a constant rate forever.

Now suppose that not only are the exogenous disturbance sequences bounded, but that after some finite date T, they are expected to be constant: $\rho_t = \rho_{LR}$ and $\bar{\imath}_t = \bar{\imath}_{LR}$ for all $t \geq T$, where the long-run values need not equal zero. We show in the Appendix that in *any* of the continuum of bounded PFE solutions, the elements of x_t converge asymptotically to long-run values

$$\pi_{LR} = \bar{\imath}_{LR} - \rho_{LR}, \qquad y_{LR} = \frac{1-\beta}{\kappa}(\bar{\imath}_{LR} - \rho_{LR}). \quad (2.17)$$

One observes that the long-run inflation rate increases one-for-one with increases in the long-run interest-rate target. Hence if we suppose that the economy must follow one or another of the PFE associated with the central bank's policy commitment, we would conclude that a lower path for the nominal interest rate must at least eventually result in a lower rate of inflation; and similarly, a higher nominal interest rate must eventually make inflation higher.

One might obtain even stronger conclusions under further assumptions about how to select a particular solution from among the set of PFE. Consider the simple case of a policy commitment under which the interest rate will be fixed at one level $\bar{\imath}_{SR}$

for all $0 \leq t < T$, and another (possibly different) level $\bar{\imath}_{LR}$ for all $t \geq T$, and let us suppose for simplicity that $\rho_t = 0$ for all t.[21] In this case, the complete set of PFE solutions (2.16) are of the form

$$\begin{aligned} x_t &= -v_1(e_1'b)\left[\frac{1-\mu_1^{T-t}}{1-\mu_1}\,\bar{\imath}_{SR} + \frac{\mu_1^{T-t}}{1-\mu_1}\,\bar{\imath}_{LR}\right] \\ &\qquad + v_2(e_2'b)\left[\frac{\mu_2^{-1}}{1-\mu_2^{-1}}\,\bar{\imath}_{SR}\right] + \chi\, v_2\mu_2^{-t} \end{aligned} \tag{2.18}$$

for all $0 \leq t \leq T$, and

$$\begin{aligned} x_t &= -v_1(e_1'b)\left[\frac{1}{1-\mu_1}\,\bar{\imath}_{\infty}\right] \\ &\qquad + v_2(e_2'b)\left[\frac{\mu_2^{T-t-1}}{1-\mu_2^{-1}}\,\bar{\imath}_{SR} + \frac{\mu_2^{-1}-\mu_2^{T-t-1}}{1-\mu_2^{-1}}\,\bar{\imath}_{LR}\right] + \chi\, v_2\mu_2^{-t} \end{aligned} \tag{2.19}$$

for all $t \geq T$.

Now suppose that we believe that there should be a unique prediction regarding the equilibrium outcomes under such a policy; then equilibrium output and inflation in period t should be given by a single-valued outcome function $x(t, T; \bar{\imath}_{SR}, \bar{\imath}_{LR})$. And suppose further that we demand that the equilibrium outcomes from any period $k > 0$ onward should also be given by *the same* outcome function, if period k is re-numbered as period 0, and all periods $t > k$ are re-numbered as $t - k$, given that the structural equations that define PFE from period k onward are of exactly the same form as those that define PFE from period 0 onward, with this re-numbering of the periods.[22] This means that the outcome function can depend only on $t - T$, rather than on the absolute magnitudes of either t or T. But the only one of the PFE given by (2.18)–(2.19) with this property is the solution with $\chi = 0$. Under this criterion, there is a

[21] Given the linearity of the model's structural equations, it is reasonable to suppose that the prediction in the case of any disturbance sequences $\{\rho_t, \bar{\imath}_t\}$ can be expressed as the sum of a predicted effect of the real disturbance $\{\rho_t\}$ (under the assumption that $\bar{\imath}_t = 0$ for all t) and a predicted effect of the monetary policy disturbance $\{\bar{\imath}_t\}$ (under the assumption that $\rho_t = 0$ for all t). The discussion in the text concerns the latter half of this problem; but similar considerations can be offered to select a particular prediction regarding the effects of alternative sequences $\{\rho_t\}$.

[22] This assumption about the form of the correct prediction is in the spirit of McCallum's (1983, 1999) "minimal-state-variable criterion." It requires that the predicted outcome in any period not depend on the number of the period, but only on its distance from the period T in which the interest rate changes, since the equilibrium conditions do not involve the former state variable.

unique PFE prediction, obtained by substituting $\chi = 0$ into equations (2.18)–(2.19). This is also the unique PFE solution selected by the "backward stability" criterion proposed by Cochrane (2015a).

Under this equilibrium selection, the perfect foresight analysis yields a "neo-Fisherian" conclusion about the effects of interest-rate policy in the short run, and not just in the long run. When we set χ to zero, the solution (2.18)–(2.19) implies that the inflation rate will equal

$$\pi_t \;=\; \lambda(t-T)\,\bar{\imath}_{SR} \;+\; (1-\lambda(t-T))\,\bar{\imath}_{LR},$$

where the sequence of weights $\{\lambda(t-T)\}$ depend only on the distance in time from the date of the policy shift. Increasing both $\bar{\imath}_{SR}$ and $\bar{\imath}_{LR}$ by the same number of percentage points is predicted to increase the inflation rate *in all periods* by exactly that same number of percentage points. Increasing only *one* of the interest rates is also predicted to increase the inflation rate both initially and later, and an increase in $\bar{\imath}_{SR}$ should immediately increase inflation by nearly as much as the increase in the interest rate, even though the increase is not expected to be permanent, if T is far enough in the future.

The conclusions that one obtains about the sign of the effects of a shift in the anticipated path of $\{\bar{\imath}_t\}$ on inflation seem then to depend crucially on the magnitude of the reaction coefficients (ϕ_π, ϕ_y), if one believes the results of the perfect foresight analysis. We shall argue however, that the conclusions of PFE analysis are misleading in the case just discussed, in which the "Taylor principle" is violated. This requires that we consider whether the PFE paths just discussed are ones that can be justified as resulting from beliefs that people would arrive at under a process of reflection, that involves a comparison between their beliefs and the outcomes that should be expected to result from such beliefs.

2.3 Reflective Equilibrium

Why should people have the particular expectations about the future that are assumed in a perfect foresight equilibrium? One answer could be that, if they expected a future path for the economy of any *other* type, action on the basis of their expectations would produce outcomes that disconfirm those expectations. One might suppose, then, that experience should sooner or later disabuse people of any expectations that are not consistent with a PFE. And if one supposes that people have

sufficient structural knowledge (including understanding of the central bank's intentions regarding future policy), one might even think that they should be able to recognize the inconsistency between their expectations and what they should expect to happen if others were also to think that way — allowing them to refine their beliefs on the basis of their understanding of how the economy works, even prior to any experience with the current economic disturbance or policy regime. Here we explore the conditions under which a PFE might arise (or under which outcomes would at least approximate a PFE) as a result of reflection of the latter sort.

Presumably the structural knowledge required for such reasoning would have to reflect previous observation of the economy; but it might reflect only experience with the effects of shocks and/or policy changes of different magnitudes and different degrees of persistence than the ones currently faced. If it were possible to acquire such knowledge of how the economy responds to shocks and to monetary policy from past experience, and then *also* to have reason to expect time paths for the current shock and current policy regime that are different from any prior experience, a process of reflection of the kind proposed here would make more sense than simply forecasting by extrapolating the past evolution of the variables that are forecasted. This kind of reasoning is particularly relevant when considering how one should expect people to respond to an announcement about future policy intentions that are historically unprecedented, as with recent experiments in "forward guidance."

We model a process of reflection by a decisionmaker (DM) who understands how the economy works — that is, who knows the correct quantitative specification of the TE relations (2.4) and (2.7) — and who also understands the policy intentions of the central bank, in the sense of knowing the policy rule (2.8) that will determine policy in all future periods. However, while the DM understands (and fully believes) the announcement of what the central bank will do, she does not know, without further reflection, what this implies about the future evolution of national income, inflation, or the resulting level of interest rates (unless the policy rule specifies a fixed interest rate).

The assumed structural knowledge can however be used to refine her expectations about the evolution of those variables. Suppose that the DM starts with some conjecture about the future evolution of the economy, which we can summarize by paths for the variables $\{e_t\}$ for each of the dates $t \geq 0$, where $t = 0$ means the date at which the economy's future evolution is being contemplated. She can then ask:

suppose that others were sophisticated enough to have exactly these expectations (on average), both now and at all of the future dates under consideration. What path for the economy should *she* expect, given her structural knowledge, under this conjecture about others' average expectations? (Note that specification of the conjecture in terms of the implied sequences $\{e_t\}$ for $t \geq 0$ gives exactly the information that is needed to answer this question, using the TE relations and the assumed path for the central-bank reaction function.)

Under such a conjecture $\{e_t\}$, the TE relations imply unique paths for the variables $\{a_t\}$, where in each period the implied a_t is given by (2.10). From these predictions the DM can infer implied paths $\{e_t^*\}$ for all $t \geq 0$, where for each date t, e_t^* are the forecasts that would be *correct* at that date if the economy evolves in the way implied by the TE relations, in the case of the average expectations $\{e_t\}$. This deduction yields an affine operator[23]

$$e^* = \Psi e$$

mapping sequences $\{e_t\}$ of conjectured expectations into sequences $\{e_t^*\}$ of correct forecasts of the same variables.[24] Note that the operator Ψ is purely forward-looking; in fact, we can write

$$e_t^* = \sum_{j=1}^{\infty} \psi_j e_{t+j} + \sum_{j=1}^{\infty} \varphi_j \omega_{t+j} \tag{2.20}$$

for all $t \geq 0$, where the sequences of matrices $\{\psi_j\}$ and $\{\varphi_j\}$ are given by

$$\psi_j \equiv (I - \Lambda)\Lambda^{j-1} M, \qquad \varphi_j \equiv (I - \Lambda)\Lambda^{j-1} m, \qquad \Lambda \equiv \begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{pmatrix}$$

for all $j \geq 1$.

We suppose (following the logic sketched above) that an awareness that the implied correct sequences e^* differ from the conjectured sequences e should constitute a

[23]Note that the definition of the operator Ψ depends on the sequences of fundamental perturbations $\{\omega_t\}$. To simplify notation, we suppress these additional arguments. We shall be interested in the application of this operator to different possible conjectured beliefs $\{e_t\}$, holding fixed the fundamentals.

[24]Note that the definition of the operator Ψ depends on the sequences of fundamental perturbations $\{\omega_t\}$. To simplify notation, we suppress these additional arguments. We shall be interested in the application of this operator to different possible conjectured beliefs $\{e_t\}$, holding fixed the fundamentals.

reason to doubt the reasonableness of expecting people to hold the conjectured beliefs. But what should one expect instead? We propose that the conjectured beliefs should be adjusted in the direction of the discrepancy between the model prediction on the basis of the conjectured beliefs and the conjectured beliefs themselves. Specifically, we consider a process of belief revision described by a differential equation

$$\dot{e}_t(n) \;=\; e_t^*(n) \;-\; e_t(n), \tag{2.21}$$

where the continuous variable $n \geq 0$ indexes how far the process of reflection has been carried forward, $e_t(n)$ is the conjecture regarding average beliefs in period t at stage n of the process, $e_t^*(n)$ is the implied correct forecast in period t if average expectations are given by the stage n conjectures, and $\dot{e}_t(n)$ is the derivative of $e_t(n)$ with respect to n.

One possible interpretation of the law of motion (2.21) for the belief-revision process would be as follows.[25] At each stage n of the process, one conjectures a particular sequence of average forecasts $\{e_t(n)\}$. But one supposes that people ought to further revise their beliefs, and considers the consequences of their revising their beliefs each time an event occurs that arrives as an independent Poisson process for each member of the population, with some fixed rate. If one supposes that each time someone revises their own expectations, they switch from whatever expectations they had held until that point, to the expectations that would be correct given the distribution of beliefs held by others *at that state of the process of belief revision,* then the rate of change of *average* beliefs will be given by (2.21); for the average of the previous period-t expectations of those revising their beliefs in any small time window will be $e_t(n)$, while the expectations that they adopt after reconsidering reconsidering their beliefs will be $e_t^*(n)$.

We suppose that the process of reflection starts from some initial "naive" conjecture about average expectations $e_t(0)$, and that the differential equations (2.21) are then integrated forward from those initial conditions. This initial conjecture might be based on the forecasts that *would* have been correct, but for the occurrence of the unusual shock and/or the change in policy that are the occasion for the process

[25]Note that this is not the *only* possible interpretation of the equation, as it specifies only a dynamic process for average beliefs, and not the heterogeneity in beliefs that may exist at each stage of the process. An alternative interpretation, in which n indexes the Poisson distribution of discrete "levels of thinking" in the population, is discussed below in section 2.4.

of reflection about what to expect in light of new circumstances. (One might suppose that but for these changes, the situation would have been a sufficiently routine one for people to have learned how to accurately forecast the economy's evolution.) The process of belief revision might be integrated forward to an arbitrary extent, but like Evans and Ramey (1992, 1995, 1998), we suppose that it would typically be terminated at some finite stage n, even if the sequences $\{e_t^*(n)\}$ still differ from $\{e_t(n)\}$.

By a *reflective equilibrium*[26] we mean a situation in which output, inflation and the nominal interest rate at some date (here numbered date 0) are determined by the TE relations, using the average subjective expectations $e_0(n)$ at the stage n at which the belief revision process is terminated.[27] We may also refer to the entire sequence of outcomes in periods $t \geq 0$ implied by the TE relations if average subjective expectations in each period are given by $e_t(n)$ as representing a reflective equilibrium of degree n. One should however only expect this entire sequence of outcomes to be realized on the assumption that the *same process of reflection* would determine beliefs and hence actions in each of the subsequent periods; this would make sense only if one supposes that the assumptions used as inputs to the process of reflection do not change in later periods in the light of additional observations, or that the process of reflection is only undertaken once. More generally, one might suppose that at each date the outcomes result from a process of reflection of the above type undertaken *at that date*, starting from an initial conjecture that may have been modified relative to the one used in previous periods; but here we consider only the beliefs resulting from a one-time process of reflection, and how similar or not these should be to PFE

[26]It may be objected that we should not speak of "equilibrium" if there remains a discrepancy between $\{e_t^*(n)\}$ and $\{e_t(n)\}$; but we use the term in the same way that Hicks and Grandmont refer to a "temporary equilibrium" (even though the assumed expectations need not be model-consistent). A reflective equilibrium is a temporary equilibrium in which the subjective expectations from which decision rules are derived are not arbitrarily specified, but instead result from the process of reflection just described.

[27]We could generalize the concept to consider possible TE outcomes resulting from distributions of subjective expectations in which different members of the population have carried forward the belief revision process to different degrees, so that there would more generally be a distribution of values of n, rather than a single value as assumed in the discussion here. Note however that, as discussed in section 2.4, the reflective equilibrium of degree n defined here can already be viewed as one in which there is a distribution of different levels of reflection across the population, with n indicating the mean level of reflection rather than a level common to all individuals.

beliefs.

A reflective equilibrium of the kind defined here might alternatively be supposed to arise from prior experience with similar shock/policy scenarios. One might suppose that the first time such a scenario occurred, people had (on average) the "naive" expectations specified by the sequence $e(0)$, and that as a result the outcome was the one implied by the TE relations (2.10) given these beliefs. The next time that people realize that a similar shock and policy response are occurring, one might suppose that (in light of the previous experience) average expectations $e(1)$[28] would instead be some convex combination of the previous expectations $e(0)$ and what turned out to be correct that time, the sequence $e^* = \Psi(e(0))$. Hence the change in expectations between the first occurrence of the scenario and the second, $e(1) - e(0)$, will be proportional to the discrepancy $\Psi(e(0)) - e(0)$ Similarly, if on the third occasion that the same scenario is expected to occur, average expectations $e(2)$ are a convex combination of average expectations $e(1)$ on the second occasion and what turned out to be correct that time, then the change $e(2) - e(1)$ will be proportional to $\Psi(e(1)) - e(1)$.

One obtains in this way an adjustment process for expectations that is essentially a discrete version of the continuous process specified in (2.21). We could on this ground be interested in whether the process (2.21), or a discretization of it, will converge eventually to a perfect foresight equilibrium, even if we do not believe that many people possess the structural knowledge reflected in the Ψ mapping. (In the adaptive learning dynamics just described, it is the working of the economy that computes $\Psi(e)$ if average expectations are e; no one in the economy need have been able to anticipate this through mental calculation.) However, our primary interest in this paper is in the question of what should happen when an unusual, perhaps wholly unprecedented, policy commitment is announced; and in such a case, our concept of a reflective equilibrium is relevant only on the assumption that at least part of the population can (at least approximately) calculate $\Psi(e)$ on the basis of their understanding of the economy. But even so, it is important to remember that (2.21) simply indicates the rate of adjustment of *average* expectations; a reflective equilibrium of degree $n > 0$ is still consistent with part of the population maintaining the "naive" expectations that they held without any consideration of what the TE

[28]Here we use $e(n)$ to mean average expectations when there have been n previous occurrences of the scenario.

relations should imply.

It should be evident that the proposed concept of reflective equilibrium will not generally lead to a unique prediction as to how the economy should evolve as a result of a specific policy commitment; the reflective equilibrium outcome will depend both on the initial expectations $e(0)$ from which the process of reflection is assumed to start, and on the stage n at which the process of reflection is assumed to terminate. Nonetheless, if the dynamics (2.21) converge globally (or at least for a large enough set of possible initial conditions) to a particular PFE, and furthermore converge rapidly enough, then a quite specific prediction will be possible under fairly robust assumptions. This is the case in which it would be justifiable to use the PFE (the *specific* PFE that represents this limit of the process of belief revision) as a prediction for what should happen under the policy commitment in question; for in this case, a reflective equilibrium will be quite similar to this PFE under a wide range of assumptions (both a wide range of assumptions about the initial beliefs $e(0)$ and a wide range of assumptions about the degree n at which the process of reflection terminates).

We show below that there exist circumstances under which PFE analysis can indeed be given a justification of this kind; in particular, we show in section 3 that when policy is expected to conform to a Taylor rule, the belief revision dynamics converge to a PFE, and more specifically to the FS-PFE defined in (2.15). This provides not only a justification for the concept of a PFE, but a definite answer to the question of which of the two-dimensional continuum of PFE solutions to (2.13) should be viewed as the model's prediction. However, in cases where the belief revision dynamics do not converge, or converge only very slowly, we argue that there is little reason to expect the outcome of a policy to be similar to the prediction of a PFE analysis.

To be sure, in such cases, the concept of reflective equilibrium will not provide a precise quantitative prediction, but only indicate a range of possibilities. But this does not mean that the predictions of a PFE analysis should be considered "as likely as anything else" to be correct. For even when the analysis of reflective equilibrium suggests only a *range* of possibilities, they may all be quite different from the conclusions that would be obtained from considering the PFE consistent with the policy in question. We show below that this is true in the case of a commitment to a fixed interest rate for a long period of time.

2.4 Related Proposals

The idea that a perfect foresight equilibrium can be obtained as the limit of an iterative process with a logic of the kind proposed above is the basis for an algorithm proposed by Fair and Taylor (1983) for numerical solution for the rational-expectations equilibrium (REE) of dynamic economic models. Essentially, their method begins with an initial conjectured sequence $e(0)$ of expectations, and computes an updated sequence of expectations $e(1) = \Psi(e(0))$ by solving the model under the conjectured expectations. This process can be repeated, resulting in a further updated sequence $e(2) = \Psi(e(1))$, and so on; the process is continued until one finds that one has a sequence of conjectured expectations that is close enough to being a fixed point of the mapping.[29] This is essentially a discrete version of the belief-revision dynamics (2.21). An important difference, of course, is that for Fair and Taylor these dynamics are simply a way for an econometrician to deduce the predictions of a model that he wishes to estimate; a failure of the dynamics to converge,[30] or to converge quickly enough, may pose a problem for the econometrician's ability to draw conclusions, but is not viewed as affecting the validity of the assumption that the data are generated in accordance with the REE of an appropriately parameterized model.

An algorithm of this kind is instead proposed as a representation of how equilibrium is actually determined in the economy, as a consequence of the calculations upon which people base their decisions, in the "calculation equilibrium" of Evans and Ramey (1992, 1995, 1998), already mentioned in section 1. Like us, Evans and

[29]The Fair-Taylor algorithm is actually more complex than this. Because they are interested in performing actual computations (with only a finite number of operations), they approximate the forward path of the economy by a sequence that extends only to some finite horizon T. But this raises the problem that it is only possible to solve a forward-looking model for endogenous variables out to date T using conjectured expectations that extend farther into the future than date T, and one cannot get such expectations by numerical solution of the model (under some previous conjecture about expectations) if one only solves the model out to date T. Their "extended path" method proposes a way to get around this problem; this complication is not needed in our exposition here, as we define our updating operation on infinite sequences, even if actual mental operations would have at best to approximate this.

[30]Fair and Taylor recognize that their algorithm need not converge. They offer a conjecture that it "will converge in the class of [RE] models for which the uniqueness conditions hold" (p. 1179) — that is, when the model has a unique forward-stable solution. In fact, that is just what we find in the case of the NK model considered here; but this determinacy condition fails in the case of an exogenous fixed path for the interest rate, as discussed in section 2.2 above.

Ramey are interested not simply in whether this process would eventually converge to a PFE, but in how quickly it converges, as they posit that the process should be terminated after only a few steps owing to "calculation costs" (that they explicitly model, unlike us). In addition to seeking to endogenize the finite degree of reflection, their approach differs from ours in considering discrete iterations of the Φ mapping, rather than a continuous belief revision process (2.21).

A related idea has also been proposed as an explanation of observed behavior in laboratory experiments with games of full information, under the name of "level-k thinking" (Stahl and Wilson, 1994, 1995; Nagel, 1995; Crawford *et al.,* 2013). This model begins by positing a "naive" form of behavior, requiring no strategic reasoning on the basis of information supplied about others' payoffs, which is taken to be the behavior of "level-0" players. "Level-1" players instead use their understanding of the game to calculate their best action on the assumption that the other players in the game think like "level-0" players; "level-2" players calculate their best action on the assumption that the other players think like "level-1" players, and so on.

Under a suitable assumption about the play of "level-0" players, the observed play of many experimental subjects in multi-player games — when the subjects are confronting a new situation (they have no experience playing the game) but have had the rules explained to them (so that they possess the structural knowledge to calculate the best response to a conjecture about others' expectations) — is found to correspond to one or another of these levels of reasoning (most commonly, levels 0, 1, 2 or 3).[31] The empirical support for this type of reasoning suggests that one should only expect an outcome similar to the Nash equilibrium prediction in cases where iteration of the best-response mapping (the analog of our Ψ mapping above) converges relatively quickly to the Nash equilibrium. (In fact, the concept has primarily been of interest

[31]Keynes (1936) famously asserted (with regard to the role of higher-order expectations in investment decisions) that few investors reason beyond "the third degree, where we devote our intelligences to anticipating what average opinion expects the average opinion to be," though "there are some, I believe, who practice the fourth, fifth and higher degrees" [p. 156]. Arad and Rubinstein (2012) report that even in the case of a fairly sophisticated population of experimental subjects, and a game which makes iterated best-response reasoning relatively natural, few if any subjects exhibit a level higher than 3. (In their preferred model of their experimental data, level-3 thinkers make up 43 percent of the sample, while the mean level of thinking is 2.2.) See Camerer *et al.* (2004) and Crawford *et al.* (2013) for reviews of other empirical evidence as to the levels of thinking observed in experimental games.

because of cases in which level-k thinking allows outcomes that remain quite different from Nash equilibrium play, for low values of k.)

Given the empirical support for level-k thinking in the experimental game theory literature, as well as the way that belief revision is modeled by Evans and Ramey, it may be wondered why we do not consider a discrete a sequence of belief revisions,

$$e_t(k) = \Psi^k(e(0)) \tag{2.22}$$

for integral levels of reasoning $k \geq 0$, instead of the continuous process (2.21). While this would lead to conclusions somewhat like those that we obtain, we prefer the continuous model of belief revision for several reasons. One is that we are concerned with *average* beliefs about *average* beliefs about ..., in an economy made up of very many individuals reflecting individually about what to do. Even if we suppose that each individual is a level-k thinker for some integral value of k, there is no reason to assume that everyone in the economy should carry the process forward for exactly the same number of stages. And a reflective equilibrium of degree n, as we have defined it, is observationally equivalent to an economy made up of level-k thinkers, under a particular population distribution of the levels of thinking.

The linear system (2.21) can be integrated forward from the given initial condition $e(0)$ to obtain

$$e(n) \;=\; \exp[n(\Psi - I)]\, e(0) \tag{2.23}$$

where I is the identity operator (mapping an infinite sequence e into itself) and the exponential of a linear operator A is defined as[32]

$$\exp[A] \;\equiv\; \sum_{k=0}^{\infty} \frac{A^k}{k!}. \tag{2.24}$$

It follows that we can write

$$e(n) \;=\; \sum_{k=0}^{\infty} s_k(n)\, e(k)$$

where

$$s_k(n) \equiv e^{-n} \frac{n^k}{k!}$$

[32]See, for example, Hirsch and Smale (1974), pp. 82-87.

for each integer $k \geq 0$, and $e(k)$ refers to the "level-k" expectations defined in (2.22). Moreover, for any n, the $\{s_k(n)\}$ are a sequence of positive weights that sum to 1, corresponding to a Poisson distribution with parameter n.

Hence a reflective equilibrium of degree n involves the same average expectations, and hence the same temporary equilibrium outcomes for output, inflation and interest rates, as a model in which fraction $s_k(n)$ of the population is made up of level-k thinkers, for each $k \geq 0$. In this interpretation, the degree of reflection n indexes a one-parameter family of possible distributions of "levels of thinking" in the sense proposed in the "level-k" literature, and the continuous variable n indicates the mean "level of thinking" in the population.[33] However, our model does not require that we assume that all (or any) members of the population have beliefs corresponding exactly to one of those in the sequence (2.22); for example, we might suppose that most (or all) people conjecture that the rest of the economy is made up of a non-degenerate distribution of different levels of thinking, as proposed by Camerer *et al.* (2004).[34]

Consideration of the continuous process (2.21) rather than a discrete sequence of progressively higher levels of thinking defined by (2.22) also avoids a technical problem with the latter progression as a way of modeling convergence to PFE. It is possible for the mapping Ψ to be such that the sequence of progressively revised beliefs $\{e(k)\}$ defined by (2.22) will fail to converge as k is made large, owing to the existence of a negative eigenvalue of absolute value greater than one, resulting in explosive oscillations.[35] Expectations of high inflation at one level of thinking result in expectations of low inflation at the next level, which result in expectations of still higher inflation at the following level, which result in expectations of still lower

[33]The use of a Poisson distribution to characterize the distribution of "levels of thinking" by a single parameter has been proposed in the experimental game theory literature by Camerer *et al.* (2004), though they also define the discrete "levels of thinking" in a different way than is standard in the "level-k" literature.

[34]Note that the interpretation of (2.21) suggested above, in the text immediately following the equation, is one in which at each stage n of the belief-revision process, not only is there a non-degenerate distribution of degrees of reflection across the population, but most members of the population have expectations that reflect an assumption that the beliefs of others involve a non-degenerate distribution of degrees of reflection. We do not present a formal analysis of this "cognitive hierarchy," as we are here concerned solely with the aggregate dynamics predicted by our model.

[35]We discuss in the Appendix how this problem can arise, for certain numerical parameter values, in the context of the model treated in this paper.

inflation at the level after that ....

In our view, the possibility of oscillations of this kind should not constitute a reason to find it unlikely that people would arrive at PFE beliefs even if able to carry out a very long chain of reasoning about others' likely beliefs. For the instability indicated by the explosive sequence requires a very rigid and implausible kind of reasoning: one must first entertain the belief that everyone else is *exactly* a level-9 thinker, and then pass from this to the conclusion that really everyone else should be *exactly* a level-10 thinker, and so on, even though the extreme conjecture at each stage of the reasoning has implications quite different from the one before. Assuming instead a continuous revision of average beliefs (which may be interpreted as a continuous shift in the population fractions that stop at different levels of thinking) avoids this possibility — though as we shall see, it still allows for belief revision dynamics that may fail to converge (for reasons that are less fragile).

Our continuous process of belief revision (2.21) is also closely related to the concept of expectational stability (or "E-stability") analyzed by Evans and Honkapohja (2001). Evans and Honkapohja classify rational-expectations equilibria as E-stable or not through an analysis of the properties of a mapping that associates with each of a class of a possible "perceived laws of motion" (on the part of the decisionmakers in some economic model) the "actual law of motion" for endogenous variables that will result from the expectations implied by that perceived law of motion. The key to their analysis is thus a mapping from a parametric specification of subjective beliefs to the corresponding specification of beliefs that would be correct if people generally act on the basis of the subjective beliefs, like our Ψ mapping. They then posit a differential equation for the adjustment of subjective beliefs (specifically, of the parameters of the "perceived law of motion") similar to (2.21), and say that an REE (a fixed point of the Ψ mapping) is "E-stable" if and only if the posited dynamics converge to it starting from arbitrary initial beliefs. The analysis of E-stability is proposed as a criterion to distinguish REE that could arise as the outcome of a learning process from ones that one should not expect to arise.

Our approach differs somewhat from theirs in that we prefer to map subjective expectations into the expectations that would be correct if people acted upon the subjective expectations, rather than mapping a subjective description of the process generating the data into the description that would be correct if people's forecasts were based on the subjective description; but while we think that our approach

has advantages,[36] it makes little difference for the conclusions obtained. The more important difference is that we parameterize beliefs in terms of sequences that describe people's beliefs about different horizons, instead of assuming that the dynamics can be described by a Markov process with only a finite number of parameters. This in turn means that our process is more easily interpreted as a process of prospective calculation by decisionmakers before they observe what actually happens, rather than a process of learning from experience. Except for this, our analysis of the convergence of reflective equilibrium to a PFE poses essentially the same question as Evans and Honkapohja consider when they determine whether such an equilibrium is E-stable.

Our convergence analysis is even more closely related to Guesnerie's (1992, 2008) consideration of the "eductive stability" of REE. Instead of assuming that only the REE of a given model (including the specification of policy) is a relevant prediction of the model, Guesnerie proposes that one should consider the entire set of outcomes that are rationalizable, in the sense that the outcome could result from optimizing behavior under some specification of expectations regarding the economy's evolution, which expectations are for outcomes that could result from optimizing behavior under still other specifications of expectations, which other expectations are for outcomes that could result from optimizing behavior, and so on.[37] The question whether a given outcome can be rationalized by progressively higher-order specifications of beliefs belonging to some admissible set $\mathcal{M}$ is essentially a question whether the outcome can be generated by beliefs that remain within the image of $\mathcal{M}$ under progressively higher-order iterations of the mapping Ψ; hence the question of "eductive stability" is essentially a question about whether the sequence defined by (2.22) converges to a given PFE for all initial beliefs $e(0)$ drawn from an admissible set. This in turn is closely related to (though not identical to) the question of convergence of the continuous process (2.21).

Like ours, Guesnerie's analysis considers whether an REE should be the outcome of a process of reflection based on knowledge of the model of the economy (including a rule that specifies policy). The most important difference from our analysis is his

[36] We need only describe subjective beliefs in terms of the evolution of two variables per period, the two summary measures of expectations that matter, rather than having to describe the evolution of the three variables per period that matter for people's decision problems.

[37] This line of analysis was originated by Phelps (1983). See Woodford (2013) for further discussion of the proposal, and its application to an NK model of the kind assumed in this paper.

consideration of belief specifications that result from repeated application of the Ψ mapping, rather than from integration of the continuous process (2.21). As discussed above, this discrete form of the belief revision process converges for a smaller range of parameter values, making "eductive stability" more elusive. But we believe that Guesnerie's criterion admits too large a set of rationalizable outcomes: it does not seem likely that people capable of a high level of reflection should expect a higher rate of inflation than the PFE rate, if this is rationalizable only by a conjecture that most other people are expecting a *lower* rate of inflation than PFE rate, which would in turn be consistent with their rationality on the supposition that most of them expect most *other* people to expect a *higher* rate than the PFE rate, and so on, with the bias changing sign in a precisely choreographed way at each higher stage.[38] Our own proposed criterion implies in such a case that if the average level of reflection is high enough, the economy's evolution should not be too different from the FS-PFE prediction.

3 Convergence of Reflective Equilibrium to Perfect Foresight: The Case of a Taylor Rule

Here we show that under some circumstances, the PFE analysis (with a correct equilibrium selection) can be justified as an approximation to a reflective equilibrium, and that (for some parameter values) the degree of reflection required to approach the PFE outcome need not even be too large. The case that we consider is that in which monetary policy is expected to be specified by a Taylor-type rule of the form (2.8), where the response coefficients ϕ_π, ϕ_y are assumed to be constant over time, though we allow a time-varying path for the intercept $\{\bar{\imath}_t\}$. The allowance for time-variation in the intercept allows us to analyze policy experiments in which there may be a commitment to conduct a "looser" policy for a specified period of time, before returning to the central bank's normal reaction function; but here the temporary loosening of policy is understood as a temporary reduction of the intercept (or a temporary increase in the implicit inflation target), while the endogenous responses

[38] Guesnerie's (2008) conclusion that the PFE is not eductively stable in the case of monetary policy specified by a Taylor rule, except for a narrow range of possible parameter values, depends on oscillating constructions of this kind.

to variations in inflation in output remain the same. We further assume that these endogenous responses satisfy (2.14), so that there is a unique bounded PFE solution in the case of any bounded sequences $\{g_t, \bar{\imath}_t\}$, given by (2.15).[39]

We assume in this section that under the reflective equilibrium dynamics, the zero lower bound never binds, so that it is in fact feasible for the central bank's interest-rate target to satisfy (2.8) at all times. This is not an "equilibrium selection" assumption, since for each value of n, reflective equilibrium is uniquely defined. It is, however, an assumption that both the disturbances to fundamentals $\{\omega_t\}$ and subjective beliefs $\{e_t(n)\}$ involve small enough departures from the long-run steady state (the values of the state variables around which we have log-linearized our model) for the interest rate implied by the TE relations (including the policy specification (2.8)) to always be non-negative.

This will be unproblematic in the case of small enough disturbances, and small enough differences between initial expectations and those consistent with the long-run steady state, *if* the belief-revision dynamics (2.21) remain forever bounded. (Since the system is linear, the bound on the distance between beliefs and steady-state beliefs will be proportional to the magnitude of the perturbations of fundamentals and of initial beliefs, so that one can ensure any desired bound — in particular, one that implies that the ZLB is never violated — by choosing a small enough bound on those perturbations.) We show below that for the kind of policy considered in this section, the belief-revision dynamics are indeed bounded for all n. Hence the analysis in this section applies as long as the shock to the economy, the policy response to it, and any associated shift in the initially conjectured expectations are all small enough departures from the long-run steady state. We defer until the next section consideration of the case in which a large shock causes the ZLB to constrain policy for some period of time.

3.1 Exponentially Convergent Belief Sequences

Our results on the convergence of reflective equilibrium as the degree of reflection increases depend on starting from an initial ("naive") conjecture that is sufficiently well-behaved as forecasts far into the future are considered. We shall say that a

[39] Note that any bounded sequence $\{g_t\}$ uniquely determines a bounded sequence $\{\rho_t\}$, given by $\rho_t = \sigma^{-1}[g_t - \beta g_{t+1}]$.

sequence $\{x_t\}$ defined for all $t \geq 0$ "converges exponentially" if there exists a finite date $\bar{T}$ (possibly far in the future) such that for all $t \geq \bar{T}$, the sequence is of the form

$$x_t \;=\; x_\infty + \sum_{k=1}^{K} a_k \lambda_k^{t-\bar{T}}, \tag{3.1}$$

where x_∞ and the $\{a_k\}$ are a finite collection of real coefficients, and the $\{\lambda_k\}$ are real numbers satisfying $|\lambda_k| < 1$. This places no restrictions on the behavior of the sequence over any finite time horizon, only that it converges to its long-run value in a sufficiently regular way. We shall similarly say that a vector sequence such as $\{e_t\}$ converges exponentially if this is true of each of the individual sequences (elements of the vector).

We shall consider only the case in which the initial belief sequence $\{e_t(0)\}$ converges exponentially. This amounts to an assumption that these "naive" beliefs are of a sufficiently simple form, as respects what is anticipated about the very distant future. Note that the TE relations (2.9)–(2.10) imply that if the sequence of fundamentals $\{\omega_t\}$ converges exponentially, and a conjecture $\{e_t\}$ regarding average subjective expectations converges exponentially as well, then the correct expectations $\{e_t^*\}$ implied by this conjecture also converge exponentially. Thus if people start from an initial conjecture about others' average expectations that converges exponentially, they should be led by reflection to beliefs that also have this property. Thus the operator Ψ maps exponentially convergent belief sequences into exponentially converging belief sequences, and any finite number of iterations will similarly lead to exponentially convergent beliefs.[40] Hence our assumption of an initial conjecture that converges exponentially does not preclude an initial conjecture that may reflect some degree of sophistication; it might, for example, be based on the paths that endogenous variables were observed to take on some previous occasion when there was a shift in fundamentals described by series that converged exponentially.

[40]The conclusion requires that the sequence of fundamentals also converges exponentially, but this is a relatively innocuous assumption. It will be satisfied, for example, if the shock (and associated policy change) that create the situation that we wish to analyze have implications after some finite horizon (possibly far in the future) that converge to long-run values with dynamics that can be described by a stable autoregressive process with real roots.

3.2 Reflection Dynamics

We now consider the adjustment of the sequence $\{e_t(n)\}$ describing subjective beliefs as the process of reflection specified by (2.21) proceeds (that is, as n increases), assuming that fundamentals $\{\omega_t\}$ converge exponentially and that the initial "naive" conjecture $\{e_t(0)\}$ converges exponentially as well. Then there exists a finite date T after which all four sequences (both elements of $\{\omega_t\}$ and both elements of $\{e_t(0)\}$) have the form (3.1). There is furthermore a finite set of growth factors $\{\lambda_k\}$ such that all four sequences can be written in the form (3.1) using the same values $\{\lambda_k\}$ for each of the series.

Thus all four sequences must belong to the linear space L, consisting of all sequences that take the form (3.1) for all $t \geq \bar{T}$, where the value of $\bar{T}$, the value of K, and the values $\{\lambda_k\}$ are part of the definition of L. Note that L is a finite-dimensional linear space (specifically, one of dimension $\bar{T} + K + 1$), the elements of which can be parameterized by specifying $\{x_t\}$ for $0 \leq t \leq \bar{T} - 1$, $\{a_k\}$ for $1 \leq k \leq K$, and x_∞. We shall similarly let $L^2 \equiv L \times L$ denote the linear space of vector sequences $\{e_t\}$ such that both elements are sequences in L.

The TE relations (2.9)–(2.10) imply that if both fundamentals $\{\omega_t\}$ and a conjecture $\{e_t\}$ about average beliefs belong to L^2, then the implied correct expectations $\{e_t^*\}$ belong to L^2 as well. The dynamics (2.21) then remain forever within the finite-dimensional linear space L^2 if one starts from an initial conjecture $\{e_t(0)\}$ in L^2. Our study of the dynamics implied by (2.21) then reduces to the study of a linear differential equation system on a finite-dimensional vector space, that we can write in the form

$$\dot{\mathbf{e}}(n) \;=\; V\,\mathbf{e}(n) \;+ W\,\boldsymbol{\omega}. \tag{3.2}$$

Here $\mathbf{e}(n)$ and $\boldsymbol{\omega}$ are vectors of length $2(\bar{T} + K + 1)$, that parameterize elements of L^2 (i.e., that specify the weights on $2(\bar{T} + K + 1)$ basis vectors for that space), and V and W are square matrices of that same dimension.

We show in the Appendix that if the central bank's reaction function satisfies the Taylor Principle (2.14), each of the $2(\bar{T} + K + 1)$ eigenvalues of the matrix V has a negative real part. This implies that V must be non-singular, and the system (3.2) has a unique fixed point, given by

$$\mathbf{e}^{PF} \;\equiv\; -V^{-1}W\,\boldsymbol{\omega}. \tag{3.3}$$

Any such fixed point must correspond to a PFE solution of the model, as defined in section 2.2 above, though the converse is not true: only PFE solutions that belong to the finite-dimensional space L^2 will be fixed points of the reduced-dimension system (2.21). The unique PFE of this kind corresponds to the FS-PFE defined by (2.15).

The general solution of the linear system of differential equations (3.2) can then be written in the form

$$\mathbf{e}(n) \;=\; \mathbf{e}^{PF} \;+\; \exp(nV)\,[\mathbf{e}(0) - \mathbf{e}^{PF}] \tag{3.4}$$

for all $n \geq 0$.[41] Furthermore, the fact that each of the eigenvalues of V has a negative real part implies that

$$\lim_{n\to\infty} \exp(nV) \;=\; 0,$$

a matrix that is zero in all its elements. This yields the following important conclusion.

Proposition 1 *Consider the case of a shock sequence $\{g_t\}$ that converges exponentially, and let the forward path of policy be specified by a sequence of reaction functions (2.8), where the coefficients (ϕ_π, ϕ_x) are constant over time and satisfy (2.14), and the sequence of perturbations $\{\bar{\imath}_t\}$ converges exponentially. Then in the case of any initial conjecture $\{e_t(0)\}$ regarding average expectations that converges exponentially, the belief revision dynamics (2.21) converge as n grows without bound to the belief sequence $\{e_t^{PF}\}$ associated with the FS-PFE.*

The implied reflective equilibrium paths for output, inflation and the nominal interest rate similarly converge to the FS-PFE paths for these variables. This means that for any $\epsilon > 0$, there exists a finite $n(\epsilon)$ such that for any degree of reflection $n > n(\epsilon)$, the reflective equilibrium value will be within a distance ϵ of the FS-PFE prediction for each of the three variables and at all horizons $t \geq 0$.

Further details of the proof are given in the Appendix.

This result has several implications. First, it shows how a PFE can arise through a process of reflection of the kind proposed in section 2.3 above. But further, it indicates that only one of the two-dimensional continuum of solutions to the difference equations (2.13) represents a PFE that can be reached in this way, at least if we accept the reasonableness of starting from an initial conjecture that is well-behaved in the

[41] See, for example, Hirsch and Smale (1974), pp. 89-97.

sense assumed in the proposition.[42] Thus it provides a justification for selecting the FS-PFE as the relevant perfect-foresight prediction of the model, if by such an exercise we understand the "perfect foresight" prediction to actually mean the limiting case of a reflective equilibrium, in which the degree of reflection is unboundedly large.

Proposition 1 also shows that the proposal to use reflective equilibrium, as defined above, as one's prediction of what should happen under a given policy need not mean that one cannot obtain predictions of any precision. In the case considered here, the reflective equilibrium predictions are quite similar, for all sufficiently large values of n, rather than depending on the precise value of n that is assumed. They are also similar (in the case of a large enough degree of reflection) regardless of the initial conjecture that is assumed, as long as the initial conjecture is not extremely distant from the beliefs associated with the long-run steady state, and the initial conjecture regarding beliefs about the distant future is well-behaved in the specified sense. Finally, in this case where the concept of a reflective equilibrium with a relatively high degree of reflection leads to a sharply-defined prediction, we see that the FS-PFE provides a useful approximation to that prediction; the accuracy of this approximation should be greater the greater the degree of reflection that one assumes.

These conclusions refer to the predictions obtained from the theory of reflective equilibrium in the case that n is "large enough"; an obvious question is how large n must be for reflective equilibrium to resemble the FS-PFE. The answer to this will depend on parameter values; but at least in some cases, the required degree of reflection may not be implausibly large. We illustrate this by considering a numerical example.

Figure 1 considers an experiment in which the intercept $\bar{\imath}_t$ is lowered for 8 quarters (periods $t = 0$ through 7 of the quarterly model), but is expected to return to its normal level from quarter 8 onward. The policy to which the central bank returns in the long run is specified in accordance with Taylor (1993): the implicit inflation target π^* is 2 percent per annum, and the reaction coefficients are $\phi_\pi = 1.5, \phi_y = 0.5/4$.[43] The model's other structural parameters are those used by Denes *et al.* (2013),

[42] This includes, for example, the "naive" hypothesis that people's expectations should be unaffected by either the shock that has occurred or the resulting change in policy. This is the specific initial hypothesis assumed in the numerical illustrations below.

[43] The division of ϕ_y by 4, relative to the value quoted by Taylor (1993), reflects the fact that periods in our model are quarters, so that i_t and π_t in (2.8) are quarterly rates rather than the annual rates used in Taylor's formula.

to show that the ZLB can produce a contraction similar in magnitude to the U.S. "Great Recession," in the case of a shock to the path of $\{g_t\}$ of suitable magnitude and persistence: $\alpha = 0.784$, $\beta = 0.997$, $\sigma^{-1} = 1.22$, and $\xi = 0.125$.[44] Among other things, these imply a long-run steady-state value for the nominal interest rate of 3.23 percent per annum.[45]

We assume that $\bar{\imath}_t$ is reduced by 0.008 (in quarterly units) for the first 8 quarters; this is the maximum size of policy shift (given the above parameters) for which the ZLB does not bind in the reflective equilibria associated with any degree of reflection $n \geq 0$.[46] In computing the reflective equilibria shown in Figure 1, we assume an initial "naive" conjecture under which expectations continue to be those that are correct in the steady state with 2 percent inflation. Finally, for simplicity we consider only a pure temporary loosening of monetary policy, not motivated by any real disturbance (so that $g_t = 0$ for all t). Because our model is linear, we can separately compute the perturbations of the steady-state paths of all variables implied by a pure monetary policy shift (assuming no real disturbance and no change in the initial conjecture), the perturbations implied by a real disturbance (assuming no change in monetary policy and no change in the initial conjecture), and the perturbations implied by a change in the initial conjecture (assuming no real disturbance or change in monetary policy), and sum these to obtain the predicted effects of a scenario under which a real disturbance provokes both a change in monetary policy and a shift in the initial conjecture.[47] In the figure, we isolate the pure effect of an announced loosening of

[44] We do not pretend to offer a quantitatively realistic analysis of alternative policies that should have been available during the Great Recession; our goal in this paper is purely to explicate the conditions under which perfect foresight analysis of monetary policy commitments makes a greater or lesser amount of sense. The parameter values proposed by Denes *et al.* are of interest as a case in which an expectation of remaining at the ZLB for several quarters has very substantial effects — and in which, more generally, monetary policy anticipations have large effects — under a rational-expectations analysis. It is in this sort of case that it matters most exactly how one models expectation determination.

[45] This means that the intercept of the central-bank reaction function assumed in the long run is smaller here than in Taylor (1993); we assume the value that (in our model) is consistent with achievement of the 2 percent inflation target in the long-run steady state.

[46] As shown in Figure 1, the shock results in a zero nominal interest rate in each of the first 8 quarters, when $n = 0$. In quarter 7, the nominal interest rate is also zero for all $n \geq 0$ (and also in the FS-PFE), since the belief-revision dynamics do not change expectations regarding any of the periods from $t = 8$ onward.

[47] Of course, in order for the ZLB not to bind in the reflective equilibria, one must bound the

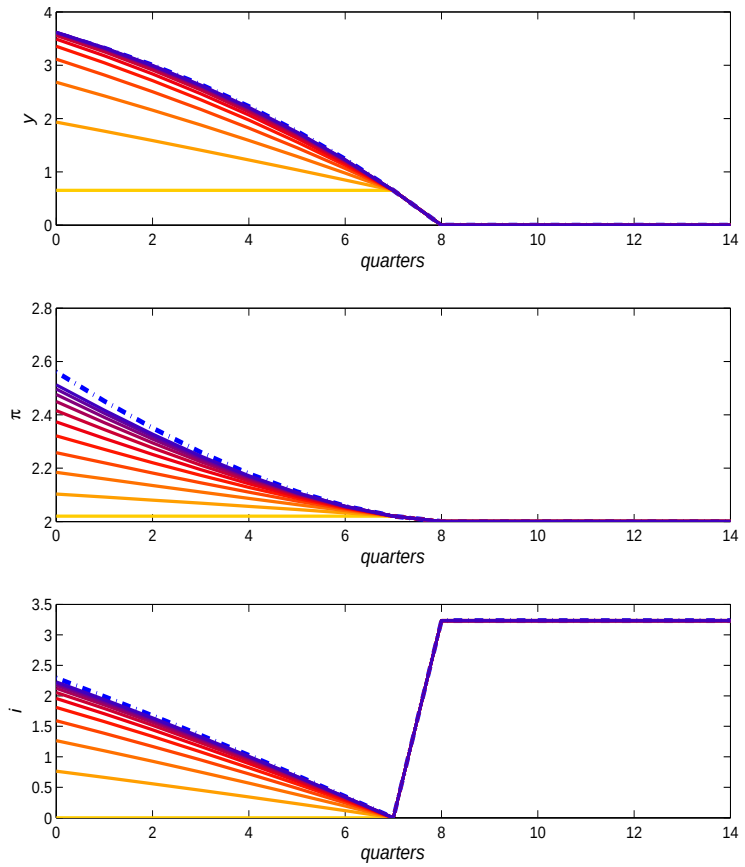

Figure 1: Reflective equilibrium outcomes for $n = 0$ through 4 (progressively darker lines) compared with the FS-PFE solution (dash-dotted line), when the Taylor-rule intercept is reduced for 8 quarters.

monetary policy, to last for a known length of time.

The three panels of the figure show the TE paths of output, inflation and the nominal interest rate,[48] in reflective equilibria corresponding to successively higher

cumulative impact of each of these three perturbations. For example, it will not be possible to loosen policy (reduce the intercept of the policy reaction function) by as much as is assumed in Figure 1 if a shock has occurred that also lowers g_t. Indeed, a sufficiently sharp temporary reduction in g_t may require the intercept of the monetary policy reaction function to be *raised* in order for the ZLB not to make implementation of the reaction function (2.8) infeasible.

[48]Here y_t is measured in percentage points of deviation from the steady-state level of output: for example, "2" means 2 percent higher than the steady-state level. The variables π_t and i_t are reported as annualized rates, and the units are again percentage points; thus "2" means two percent per annum. Note that in this and all later figures, y_t is reported as a log deviation from steady-state

degrees of reflection. The lightest of the solid lines (most yellow, if viewed in color) corresponds to $n = 0$; these are the outcomes that are expected to occur under the "naive" conjecture about average expectations (namely, that these do not take account of the policy shift at all), but taking account of the announced change in the central bank's behavior in the TE analysis under those expectations. (Thus the $n = 0$ paths do not represent the naive beliefs, but rather the paths that it would be correct to expect, if people on average hold the naive beliefs.) The relatively aggressive reduction in the interest rate has some stimulative effect on output even in the absence of any change in expectations, but this effect is the same in each of the first 8 quarters; the effect of the loosening of policy is the same, regardless of the number of additional quarters for which the loose policy will continue.

As n increases, the effects on output and inflation become greater in quarters zero through 6; and the extent to which this is so is greater, the larger the number of quarters for which the looser policy is expected to continue. There are no changes in the expected paths from quarter 8 onward, as n increases; this is because we have assumed reversion to the long-run steady-state policy in quarter 8, and the initial "naive" conjecture already corresponds to a PFE from quarter 8 onward, so that beliefs do not change as n increases.[49] There are similarly no changes in the expected outcomes in quarter 7, because quarter 7 *expectations* about later quarters do not change, in the absence of any outcomes different from those expected in any of those later quarters. However, the fact that outcomes are different in quarter 7 and earlier than those anticipated under the "naive" expectations causes beliefs to be revised in quarters 6 and earlier. As expectations shift toward expecting higher output and inflation in one or more later periods, the TE levels of output and inflation in the earlier quarters increase (and the nominal interest rate increases as well, through an endogenous policy reaction). This effect is greater the larger the number of future quarters about which expectations of output and inflation are revised upward.

The progressively darker solid lines in the figure plot the reflective equilibrium outcomes for degrees of reflection $n = 0, 0.4, 0.8$, and so on up to $n = 4.0$. The FS-PFE paths are also shown by dark dash-dotted lines. One sees from the figure that

output, as in the model equations, but π_t and i_t in the figures are shown in absolute terms, not as deviations from the steady-state values of these variables.

[49]Because the model is purely forward-looking, revisions of expectations about *earlier* periods have no effects on equilibrium determination from period 8 onward.

the reflective equilibrium paths converge to the FS-PFE solution as n increases, in accordance with Proposition 1. Moreover, the convergence is relatively fast, for this kind of policy experiment. Already when $n = 2$, the predicted reflective equilibrium responses for both output and inflation differ from the PFE responses by less than 10 percent (in fact, by less than 7.5 percent) in any quarter. This means that if the average member of the population is expected to be capable of iterating the Ψ mapping at least twice,[50] one should predict outcomes approximately the size of the PFE outcomes. When $n = 4$, the reflective equilibrium output responses differ from the PFE outcomes by only 1 percent or less, and except in quarter zero (when the discrepancy is closer to 2 percent), the same is true of the inflation responses.

Higher degrees of reflection would only make the FS-PFE prediction even more accurate. This provides a good example of the kind of situation in which, in our view, a perfect foresight equilibrium analysis of the effects of a monetary policy commitment can make sense. Note that is specifically the FS-PFE, rather than any other solution to the difference equations (2.13), that provides a good approximation to a reflective equilibrium (as long as n is not extremely low). This provides an answer to the question raised by Cochrane (2011) about the justification of appealing to the FS-PFE in monetary policy analysis.

3.3 Effects of a Policy Change Far in the Future

The paradox posed in section 1 involves arguments about the effects of an expectation that policy will be changed *permanently,* rather than for only a few quarters as in Figure 1, and questions about how much it can matter what is assumed about policy extremely far in the future. Here we consider these issues in the case of the class of policies discussed above, where (temporary or permanent) policy changes are understood simply to involve changes in the intercept of the monetary policy reaction

[50]Here it is worth recalling that Arad and Rubinstein (2012) find that their subjects have a mean "level of thinking" of 2.2. Camerer *et al.* (2004), however, conclude that "an average of 1.5 steps [of iterated best response] fits data from many games" [p. 861]. It should be noted that these experimental results relate to subjects' play in one-shot games, where the strategic considerations have been explained to the players, but they have no *experience* on the basis of which to calculate their best action. One might expect that the realistic mean (effective) degree of reflection n will be higher in cases where people have some degree of prior experience with the policy regime in question, as discussed further in section 3.4.

function.

For the sake of specificity, we consider the following special class of policy experiments. Suppose that $\bar{\imath}_t$ is expected to take one value ($\bar{\imath}_{SR}$) for all $t < T$, and another value ($\bar{\imath}_{LR}$) for all $t \geq T$. (The policy experiment considered in Figure 1 is one example of a policy in this class, with $\bar{\imath}_{SR} < 0$, $\bar{\imath}_{LR} = 0$, and $T = 8$. In this more general discussion, we again assume that both $\bar{\imath}_{SR}$ and $\bar{\imath}_{LR}$ are high enough that the ZLB never binds.) How does the effect of such a policy commitment vary depending on the choice of the horizon T? In particular, should the effect be similar for all large enough values of T?

As in the case considered in Figure 1, we consider the effects of a pure policy change, assuming $g_t = 0$ for all t and an initial "naive" conjecture in which average expectations are consistent with the steady state in which the inflation target π^* is achieved at all times.[51] Because our model is purely forward-looking, and $\bar{\imath}_t, g_t$, and $e_t(0)$ are each the same for all $t \geq T$, it is easy to see that the belief-revision dynamics (2.21) result in $e_t(n)$ having the same value for all $t \geq T$. Let this value be denoted $e_{LR}(n)$. We see that it must evolve according to

$$\dot{e}_{LR} \;=\; [M - I]\, e_{LR} \;+\; m_2\, \bar{\imath}_{LR} \tag{3.5}$$

starting from the initial condition $e_{LR}(0) = 0$, where m_2 is the second column of the matrix m in (2.10).

Let us suppose that the quantity on the left-hand side of (2.14) is not exactly equal to 1;[52] in this case we can show (see the Appendix) that $M - I$ is non-singular. The solution to (3.5) is then easily seen to be[53]

$$e_{LR}(n) \;=\; [I \;-\; \exp[n(M - I)]]\;\; \bar{e}^{PF}_{LR} \tag{3.6}$$

for all $n \geq 0$, where

$$\bar{e}^{PF}_{LR} \;\equiv\; [I - M]^{-1}\, m_2\, \bar{\imath}_{LR} \tag{3.7}$$

[51] As in the case discussed above, we can determine the effect of varying the length T of the policy commitment in the case of a given real disturbance (represented by a sequence $\{g_t\}$) by summing the effect of the pure policy change (computed here as a function of T) and the effect of the real disturbance in the absence of any policy change (which will be independent of T).

[52] Note that this condition is satisfied by generic reaction functions of the form (2.8) whether the Taylor Principle is satisfied or not. Hence we do not discuss the knife-edge case in which $M - I$ is singular, though our methods can easily be applied to that case as well.

[53] See, for example, Hirsch and Smale (1974), p. 90.

is the unique rest point of the dynamics (3.5).

Note that $\bar{e}_{LR}^{PF}$ is also the stationary vector of average expectations associated with the unique PFE steady state, in the case that the policy $\bar{\imath}_t = \bar{\imath}_{LR}$ is expected to be maintained forever. If the reaction coefficients (ϕ_π, ϕ_y) satisfy the Taylor Principle (2.14), then (as shown in the Appendix) both eigenvalues of $M - I$ have negative real part, and

$$\lim_{n\to\infty} \exp[n(M-I)] \;=\; 0. \tag{3.8}$$

It then follows from (3.6) that

$$\lim_{n\to\infty} e_{LR}(n) \;=\; \bar{e}_{LR}^{PF},$$

so that the reflective equilibrium in any period $t \geq T$ converges to the PFE steady state associated with the long-run policy (which is also the FS-PFE solution for this policy). This is of course as we should expect from Proposition 1.

We turn now to the characterization of reflective equilibrium in periods $t < T$. The forward-looking structure of the model similarly implies that the solution for $e_t(n)$ depends only on how many periods prior to period T the period t is, and not on the dates of either t or T. If we adopt the alternative numbering scheme $\tau \equiv T - t$ (i.e., we number periods according to the number remaining until the shift to the long-run policy), then the solution for $e_\tau(n)$ for any $\tau \geq 1$ will be independent of T. Moreover, in terms of this notation, the belief-revision dynamics (2.21) can be written in the form

$$\dot{e}_\tau(n) \;=\; -e_\tau(n) \;+\; \sum_{j=1}^{\tau-1}[\psi_j e_{\tau-j}(n) + \varphi_{j2}\bar{\imath}_{SR}] \;+\; \sum_{j=\tau}^{\infty}[\psi_j e_{LR}(n) + \varphi_{j2}\bar{\imath}_{LR}]$$

for each $\tau \geq 1$, where φ_{j2} is the second column of the matrix φ_j. These dynamics can equivalently be written in the form

$$\begin{aligned}\dot{e}_\tau(n) \;=\; -e_\tau(n) \;&+\; (I-\Lambda)\sum_{j=1}^{\tau-1}\Lambda^{j-1}\left[Me_{\tau-j}(n) \;+\; m_2\bar{\imath}_{SR}\right] \\ &+\; \Lambda^{\tau-1}\left[Me_{LR}(n) \;+\; m_2\bar{\imath}_{LR}\right],\end{aligned} \tag{3.9}$$

and integrated forward from the initial conditions $e_\tau(0) = 0$ for all $\tau \geq 1$, using solution (3.6) for $e_{LR}(n)$.

We observe that for $\tau = 1$, the linear differential equation (3.9) can be solved uniquely for the function $e_1(n)$, given that $e_{LR}(n)$ is already known. Then the equation for $\tau = 2$ can be solved uniquely for the function $e_2(n)$, given that $e_1(n)$ and $e_{LR}(n)$ are already known; and proceeding recursively in this way, one can solve uniquely for the $\{e_\tau(n)\}$ for all values of τ up to any given bound T (corresponding to the initial period $t = 0$). In this way, we obtain a unique solution for $e_t(n)$ for all $t \geq 0$.

Note further that considering how $e_t(n)$ changes (for any fixed t) as T is increased is equivalent to considering how the solution to the system of differential equations (3.9) changes for progressively larger values of τ. In particular, the behavior of $e_t(n)$ as T is made unboundedly large can be determined by calculating the behavior of the solution to the system (3.9) as $\tau \to \infty$. This yields the following simple result.

Proposition 2 *Consider the case in which $g_t = 0$ for all t, and let the forward path of policy be specified by a sequence of reaction functions (2.8), where the coefficients (ϕ_π, ϕ_x) are constant over time and such that the left-hand side of (2.14) is non-zero, and suppose that $\bar{\imath}_t = \bar{\imath}_{SR}$ for all $t < T$ while $\bar{\imath}_t = \bar{\imath}_{LR}$ for all $t \geq T$. Then if the initial conjecture is given by $e_t(0) = 0$ for all t, the reflective equilibrium beliefs $\{e_t(n)\}$ for any degree of reflection n converge to a well-defined limiting value*

$$e_{SR}(n) \equiv \lim_{T \to \infty} e_t(n)$$

that is independent of t, and this limit is given by

$$e_{SR}(n) = [I - \exp[n(M - I)]] \; \bar{e}^{PF}_{SR}, \tag{3.10}$$

where

$$\bar{e}^{PF}_{SR} \equiv [I - M]^{-1} m_2 \bar{\imath}_{SR}. \tag{3.11}$$

The reflective equilibrium outcomes for output, inflation and the nominal interest rate then converge as well as T is made large, to the values obtained by substituting the beliefs $e_{SR}(n)$ into the TE relations (2.10) and the reaction function (2.8).

The proof is given in the Appendix. This result implies that our concept of reflective equilibrium, for any given degree of reflection n, has the intuitively appealing property that a commitment to follow a given policy (a given intercept for the reaction

function, or a given implicit inflation target) for a time horizon T has similar consequences for all large enough values of T; moreover, for any large enough value of T, the policy that is expected to be followed after date T has little effect on equilibrium outcomes. Comparison of expressions (3.10)–(3.11) with (3.6)–(3.7) also shows that the predicted outcomes in the case of any long enough horizon T for maintenance of the "temporary" policy are close to the predicted outcomes (under a reflective equilibrium with the same degree of reflection n) in the case that the policy is expected to be permanent. In the case of policies in the class considered here, there is no relevant difference between a commitment to a given reaction function for a long but finite time and a commitment to follow the rule forever.

Next, we consider how the reflective equilibrium prediction in the case of a long horizon T changes as the degree of reflection n increases. If the coefficients (ϕ_π, ϕ_y) satisfy the Taylor Principle (2.14), then (3.8) implies that as n is made large,

$$\lim_{n\to\infty} e_{SR}(n) \;=\; \bar{e}_{SR}^{PF}.$$

Moreover, the beliefs $\bar{e}_{SR}^{PF}$ defined in (3.11) are simply the steady-state PFE beliefs (or FS-PFE beliefs) in the case of a permanent commitment to the reaction function (2.8) with $\bar{\imath}_t = \bar{\imath}_{SR}$. Thus we obtain the following result.

Proposition 3 *Suppose that in addition to the hypotheses of Proposition 2, the coefficients (ϕ_π, ϕ_y) satisfy the Taylor Principle (2.14). Then the limits*

$$\lim_{n\to\infty} \lim_{T\to\infty} e_t(n) \;=\; \lim_{n\to\infty} e_{SR}(n) \;=\; \bar{e}_{SR}^{PF}$$

and

$$\lim_{T\to\infty} \lim_{n\to\infty} e_t(n) \;=\; \lim_{T\to\infty} e_t^{PF} \;=\; \bar{e}_{SR}^{PF}$$

are well-defined and equal to one another. Moreover, both are independent of t, and equal to the FS-PFE expectations in the case of a permanent commitment to the reaction function (2.8) with $\bar{\imath}_t = \bar{\imath}_{SR}$.

Proposition 3 identifies a case in which the thought experiment of considering the PFE consistent with a permanent commitment to a given policy rule does not lead to paradoxical conclusions. Not only does the question have a unique, well-behaved answer (if one selects the FS-PFE solution, as is conventional in the NK literature), but this answer provides a good approximation to the reflective equilibrium outcome

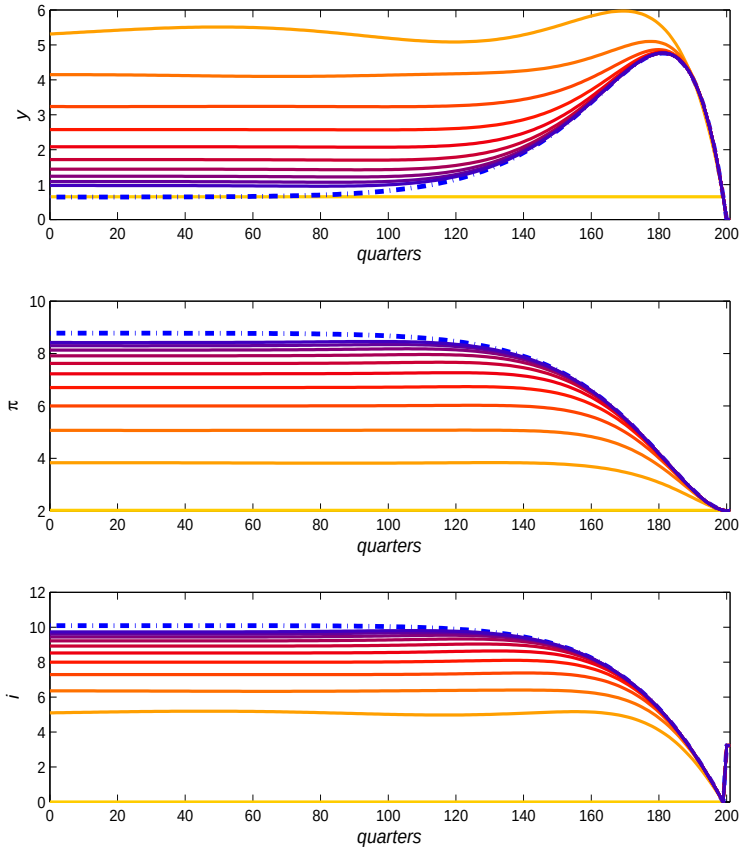

Figure 2: Reflective equilibrium outcomes for $n = 0$ through 20 (progressively darker lines) compared with the FS-PFE solution (dash-dotted line), when the Taylor-rule intercept is reduced for 200 quarters.

in the case of any large enough degree of reflection n *and* any long enough horizon T for maintenance of the policy.

Figure 2 provides a numerical illustration of these results. The policy experiment is the same as in Figure 1, as are the assumed numerical parameter values, except that in Figure 2 the commitment to the intercept $\bar{\imath} < 0$ is expected to last for 50 years. (This is not forever, but it should already be evident from the figure that further increases in the length of the commitment will make little difference in the predicted outcomes over the first decade or two of the commitment to looser policy; consideration of a finite value of T makes it still possible to show how the reflective equilibrium outcomes change for smaller values of τ, so that the results for all values of

T of 50 years or less can be shown in a single figure.) Again the reflective equilibrium paths are shown for progressively higher values of n.[54] The figure shows not only the convergence of the reflective equilibrium outcomes for all three variables as T is made large, for each of the possible values of n, but also the convergence of reflective equilibrium to the FS-PFE predictions, for each of the possible values of τ (and hence for each possible value of T). Not only is $e_\tau(n)$ close to e_τ^{PF} for all large enough n in the case of any single value of τ, but there exists a value of n for which $e_\tau(n)$ is close to e_τ^{PF} for *all* τ (and hence for all possible commitment lengths T).

While a calculation of the FS-PFE implied by a permanent commitment to a Taylor rule clearly represents a meaningful limiting case, one can not necessarily conclude that it should provide a good quantitative prediction about the effects of a policy change that is expected to be long-lasting. Figure 2 shows that (for the parameter values assumed) the FS-PFE provides a good approximation to the reflective equilibrium outcome if the degree of reflection is on the order of $n = 20$ (or even higher values that are not shown); but this would involve a degree of reflection that seems fairly unrealistic, if it is taken to represent a purely prospective calculation on the part of people who have learned about an announced policy change, but not yet had occasion to observe what actually happens under the new regime. If the degree of reflection equals only $n = 2$,[55] for example, then if the commitment is to change policy for only two years (as in Figure 1), the reflective equilibrium outcomes are not too different from the FS-PFE predictions; but when the new policy is expected to last for decades (as in Figure 2), the predicted outcomes are quite different, even if the responses to the policy change under reflective equilibrium have the same sign as the FS-PFE predictions. (The output increase predicted by the reflective equilibrium is many times larger than the FS-PFE prediction, while the inflation increase is only a fraction of the FS-PFE prediction.)

This illustrates a general point: in judging the practical relevance of the PFE prediction, it matters not only whether reflective equilibrium should converge to the PFE as n is made large enough, but also how *quickly* such convergence should occur. The speed of convergence is not too great an issue in the case of a commitment to a

[54] Again, the movement from lighter to darker lines corresponds to increasing n. The lines shown in the figure correspond to the values $n = 0, 2, 4$, and so on up to $n = 20$.

[55] In the figure, this is the line for which the output response is largest, while the inflation and interest-rate responses are the second-lowest.

new policy to be maintained for only a few quarters, when the new policy is a Taylor rule, and when the temporary policy is to be followed by a reversion to a policy regime that the public already understands well on the basis of past experience (the experiment considered in Figure 1). It is a bigger issue, however, in the case of a commitment to a new permanent (or at any rate long-lasting) regime that differs non-trivially from past policy (for example, adoption of a new inflation target that is announced as a permanent change), even when both the old and the new policies conform to the Taylor principle. And, as we show in the next section, it is a still larger issue in the case of a temporary regime under which the Taylor principle is not expected to be satisfied, as in the case of a commitment to a fixed interest rate for a significant period of time.

3.4 The Fisher Equation and Long-Lasting Shifts in Policy

We can now address a question posed in the title of our paper: what should happen if people come to expect (whether as a result of a central-bank announcement, or on the basis of experience) that a "loose" monetary policy will be maintained for several more years? Should such a shift in understanding of the outlook for future policy be inflationary, or can it be deflationary? If policy is expected to follow a Taylor rule, and "looser policy" means a lower intercept in that rule (and thus a lower nominal interest rate *for any given outcomes* for inflation and output, but not necessarily a lower nominal interest given the endogenous effects on inflation and output), then we can answer the question using the results above.

Our results show that in our model, an expectation that the reaction-function intercept will be kept lower than usual for the next several years should lead to *higher* inflation and output, regardless of the degree of reflection, and regardless of the length of time for which the looser policy is expected to be maintained. (In this respect, the FS-PFE solution gives the correct answer, at least as regards the sign of the effects.) If the loosening of policy is expected to be sufficiently transitory (though it may last for some years), as in Figure 1, then the policy change will be associated with a temporarily lower nominal interest rate, regardless of the degree of reflection. But if the shift in the policy rule is expected to last for a sufficiently long (though possibly finite) period of time, the higher inflation rate and output will be associated with a *higher* nominal interest rate, despite the reduction in the intercept

of the central bank's reaction function, except in the case of a degree of reflection that is very low. (Figure 2 shows that when $n = 2$ or greater, even a commitment to maintain looser policy for five years results in a higher nominal interest rate than the steady-state level that would represent the reflective equilibrium in the absence of a policy change.[56] Longer commitments would result in even greater increases in the nominal interest rate.)

In the case of a *permanent* increase in the inflation target, the FS-PFE prediction is that the nominal interest rate should increase one-for-one with the increase in inflation (which increases by exactly the increase in the target); the relationship between the permanent change in the inflation rate and the permanent change in the level of nominal interest rates satisfies the Fisher equation. In the case of a reflective equilibrium with only a finite degree of reflection n, the Fisher equation need *not* be satisfied (it depends on expectations being correct, at least on average), though it will hold *approximately*, if n is large. But even for modest values of n, a permanent increase in the inflation target, which permanently raises inflation, is likely to be associated with an increase, rather than a decrease, in nominal interest rates.

These results do not involve any discontinuity in the predicted effects of a policy change as one passes from the case of a long-lasting (but still temporary) change to the case of a permanent change in policy; Figure 2 shows how the predicted effects on output, inflation and the nominal interest rate all vary continuously as τ increases (and hence as T increases, for a fixed value of t). Nor do they involve any failure of the conventional expectation that *reducing* the intercept of the interest-rate reaction function represents a more expansionary (and more inflationary) policy, regardless of the length of time that the policy shift is expected to last. They do, however, indicate that the *change in the nominal interest rate* is not necessarily a good measure of the degree to which policy is loosened, as a shift *down* in the reaction function may be associated with an *increase* in the nominal interest rate. Indeed, this is almost certainly what should be observed, in the case of a sufficiently long-lasting change in policy.

Rather than supposing that the degree of reflection n is fixed, as in our formal analysis above, it is plausible to suppose that in the case of a long-lasting shift in policy, average expectations at the time of the initial announcement of the novel

[56] The effect of a commitment that lasts for $T = 20$ quarters can be read off from the figure by observing the predictions for quarter 180, which corresponds to $\tau = 20$.

policy will correspond to a relatively low value of n, but that over time the value of n should increase. The reason is not simply that people would have more time to think through the implications of the new policy regime, but (more importantly) that observation of economic outcomes under the new regime should lead people to adjust their expectations in the same direction as is implied by an increase in n.[57]

Consider, for simplicity, the case of a permanent shift in the intercept $\bar{\imath}_{LR}$, and suppose that the initial conjecture $e(0)$ is that expectations do not change at all ($e_t(0) = 0$ for all t). The reflective equilibrium for some low value of n involves $e_t(n) = e_{LR}(n)$ for all t, where $e_{LR}(n)$ is defined in (3.6). This will imply constant levels of output, inflation and the nominal interest rate corresponding to that value of n (the values that can be read off from the extreme left points of the responses shown in Figure 2). But, given these constant levels of output, inflation and interest rates, the correct expectations $e_t^*(n)$ will also be the same for all t, but different from average expectations $e_{LR}(n)$. And observing actual output, inflation and interest rates for even a few periods should indicate the direction in which outcomes under the new regime are different from those that have been expected on average.

If we suppose that as a result, people's expectations (at least on average) should shift in the direction of the discrepancy — specifically, that e_t continues to be the same for all future dates t, but that the constant value changes in proportion to the constant difference

$$e^*(n) \; - \; e_{LR}(n) \; = \; [M - I]\, e_{LR} \; + \; m_2\, \bar{\imath}_{LR}$$

— then the new time-invariant value for e_t should correspond to $e_{LR}(n)$ for a somewhat higher value of n, given that the evolution of $e_{LR}(n)$ in response to increases in n is given by (3.5). But in subsequent periods, observation of the outcomes resulting from expectations $e_{LR}(n)$ with this higher value of n should lead expectations to be

[57] In the experimental game theory literature, it is often observed that when subjects get to play a given game repeatedly, observed play deviates much less from the Nash equilibrium prediction after a few repetitions (see, e.g., Nagel, 1995). In the games in question, this is what the theory of reflective equilibrium would predict, with a fixed initial conjecture, if the average level of reflection n were to increase on each repetition. However, it could also occur without any increase in the average level of reflection, if one supposes that the initial conjecture changes on each repetition, being determined by average behavior on the previous instance; this is essentially the interpretation of her data proposed by Nagel. Stahl (1996) interprets the same data in terms of an alternative learning model, in which there is an increase over time in players' "level of thinking."

revised in a way that corresponds to a still higher value of n, and so on indefinitely if the new regime continues without further changes.

In this way, one might expect to observe over time, not the reflective equilibrium (as defined above) corresponding to a single value of n, but rather a progression from lower to steadily higher values of n. If the initial level of reflection when the policy shift is announced is quite low, then even a permanent reduction in the reaction-function intercept might initially be associated with a decrease, rather than an increase, in the nominal interest rate (as shown, for example, in Figure 2 for the case $n = 0$). However, observation of the outcomes produced by the new policy (which differ from average expectations) should cause n to increase; and at first, this should result in increases in output, inflation *and* the nominal interest rate required by the new reaction function (as shown in Figure 2 by the difference between the responses for $n = 2$ compared to those for $n = 0$).

As n increases still further (as it should with sufficient experience of the new policy), the output effect of the policy change should *decrease,* while inflation and the nominal interest rate continue to increase (as shown in Figure 2 by the movement from $n = 2$ to the cases $n = 4, n = 6$, and so on). Hence such a permanent shift in policy could be associated with an initial decrease in the nominal interest rate, though the nominal interest rate should eventually (and permanently) be increased. The policy shift should increase both output and inflation, but if the effective degree of reflection increases with experience, one would observe a much stronger output effect in the beginning, while the eventual effect would be a permanent increase in inflation and the nominal interest rate while most of the output effect would prove temporary.[58]

Thus a permanent (or long-lasting) "loosening" of policy, in the sense of a reduction of the reaction-function intercept, need not mean permanently lower nominal interest rates. (Indeed, if the change in policy is immediately understood to be permanent, and a sufficiently large part of the population engages in a sufficient degree of reflection, the period for which the nominal interest rate is reduced might not be very long.) But this doesn't mean that a central bank *couldn't* decide to maintain

[58]If one further supposes that the automatic rate of price increase π^* between occasions on which prices are re-optimized would eventually increase in the case of a permanently higher inflation rate, then the output effect would eventually disappear altogether. This is not seen in Figure 2 even as $T \to \infty$ and $n \to \infty$, because of the assumption of indexation at the fixed rate π^*.

a lower nominal interest-rate target for many years, and that it could not credibly announce an intention to do so. However, such a policy (or such an understanding of policy) is *not* equivalent to any particular degree of adjustment of the intercept of a Taylor-type rule, and our results above need not apply. We turn next to an analysis of reflective equilibrium in this alternative case.

4 Consequences of a Temporarily Fixed Nominal Interest Rate

We now consider the case in which it comes to be understood (either as a result of a shock, or a policy announcement) that the nominal interest rate will be fixed at some level $\bar{\imath}_{SR}$ up to some date T, while it will again be determined by the "normal" central bank reaction function from date T onward. (The latter policy is assumed to be a rule of the form (2.8), in which the response coefficients satisfy the Taylor Principle (2.14), and the intercept is consistent with the inflation target π^*.) There are various reasons for interest in this case. First, a real disturbance may create a situation in which the interest rate prescribed by the Taylor rule violates the ZLB for some time; in such a case, it may be reasonable to suppose that the central bank will set the nominal interest rate at the lowest possible rate, regardless of the exact outcomes for output and inflation, as long as the situation persists, but return to implementation of its normal reaction function once this is feasible. And second, a central bank may commit itself to maintain the nominal interest rate at its lower bound for a specific period of time, even if this is lower than the rate that the Taylor rule would prescribe. The "date-based forward guidance" provided by several central banks in the aftermath of the global financial crisis arguably involved commitments of this kind; and while no explicit promises were made about how policy would be conducted *beyond* the horizons in question, one might suppose that people would expect the central bank to revert to its usual approach to policy once there ceased to be any explicit commitment to behave otherwise. We are interested in the extent to which such a temporary change in policy should have effects similar or different from the effects of a temporary shift in the intercept of the monetary policy reaction function, analyzed above.

We are interested in two kinds of questions about the effects of such policies. One

is what the effect should be of changing $\bar{\imath}_{SR}$, taking the horizon T as given (perhaps by the expected persistence of an exogenous real disturbance). While there might seem to be no room to vary the short-run level of the interest rate, if we imagine a case in which it is already at the ZLB, it would even in that case always be possible to commit to a *higher* (though still fixed) interest-rate target, and some have suggested that (at least when the situation of being constrained by the ZLB persists for a long enough time) it might actually be expansionary to do so. A second question is the effect of changing T, the length of time that the interest rate is held fixed, taking as given the time path of the real disturbance. To what extent can a commitment to keep the interest rate low for a longer time substitute for an ability to cut rates more sharply right away (which may be infeasible due to the ZLB)?

4.1 Convergence to Perfect-Foresight Equilibrium

We first consider whether reflective equilibrium converges to a PFE again in this case, as n grows, and if so to which of the possible PFE paths. The question of equilibrium selection is of particular interest in this policy experiment, since here, unlike the case considered in section 3, the "backward stability" selection criterion proposed by Cochrane (2015a) would imply a different solution than the conventional "forward stability" (or local determinacy) criterion.[59]

Because of the forward-looking character of our model, the determination of reflective equilibrium from period T onward depends only on the specification of policy from period T onward. Since we again assume a reaction function that satisfies the Taylor Principle over this period, the results of section 3 continue to apply; specifically, Proposition 1 implies that in the case of any initial conjecture that converges exponentially, reflective equilibrium outcomes will converge to the unique FS-PFE outcomes as n increases. If we suppose that $g_t = 0$ for all $t \geq T$, this means that the reflective equilibrium outcomes for all $t \geq T$ will converge to the steady state consistent with the inflation target π^*. Note that this simple result already tells us that reflective equilibrium *cannot* generally converge to the "backward stable" solution proposed by Cochrane (2015a), as this does not generally imply that the long-run steady state is reached from date T onward. Instead, if reflective equilibrium converges to *any* PFE, it can only converge to the FS-PFE, which does imply steady-state

[59] See the discussion of this point by Cochrane (2015a), sec. 4.1.

outcomes from date T onward in the case just discussed.

The analysis of convergence prior to date T requires an extension of our previous result, because now we assume that the response coefficients (ϕ_π, ϕ_y) differ before and after date T. Nonetheless, as shown in the Appendix, the methods used to prove Proposition 1 can be extended to establish convergence in this case as well.

Proposition 4 *Consider the case of a shock sequence $\{g_t\}$ that converges exponentially, and let the forward path of policy be specified by a fixed interest rate $\bar{\imath}_{SR}$ for all $0 \leq t < T$, but by a reaction function of the form (2.8) for all $t \geq T$, where the coefficients (ϕ_π, ϕ_x) of the latter function satisfy (2.14), and the intercept is consistent with the inflation target π^*. Then in the case of any initial conjecture $\{e_t(0)\}$ regarding average expectations that converges exponentially, the belief revision dynamics (2.21) converge as n grows without bound to the belief sequence $\{e_t^{PF}\}$ associated with the FS-PFE.*

The implied reflective equilibrium paths for output, inflation and the nominal interest rate similarly converge to the FS-PFE paths for these variables. This means that for any $\epsilon > 0$, there exists a finite $n(\epsilon)$ such that for any degree of reflection $n > n(\epsilon)$, the reflective equilibrium value will be within a distance ϵ of the FS-PFE prediction for each of the three variables and at all horizons $t \geq 0$.

Figure 3 provides a numerical illustration of this result. The model parameters are as in the previous numerical examples, and for simplicity we again show the effects of a pure shift in monetary policy, assuming $g_t = 0$ for all t and an initial conjecture under which $e_t(0) = 0$ for all t. As in Figure 1, it is again assumed that monetary policy is expected to depart from the "normal" Taylor rule for 8 quarters, and then to revert to the "normal" reaction function thereafter. The only difference is that in Figure 3 it is assumed that the nominal interest rate is fixed at zero for the first 8 quarters.

For the case $n = 0$ (the lightest of the lines in the figure), the responses are identical to those in Figure 1: the two shifts in monetary policy have been chosen to lower the nominal interest rate to the same extent (i.e., to zero), in the absence of any change in average expectations. For higher values of n, the effects of the policy change are qualitatively similar to those in Figure 1, but not exactly the same: the output and inflation increases are somewhat larger when the interest rate is expected

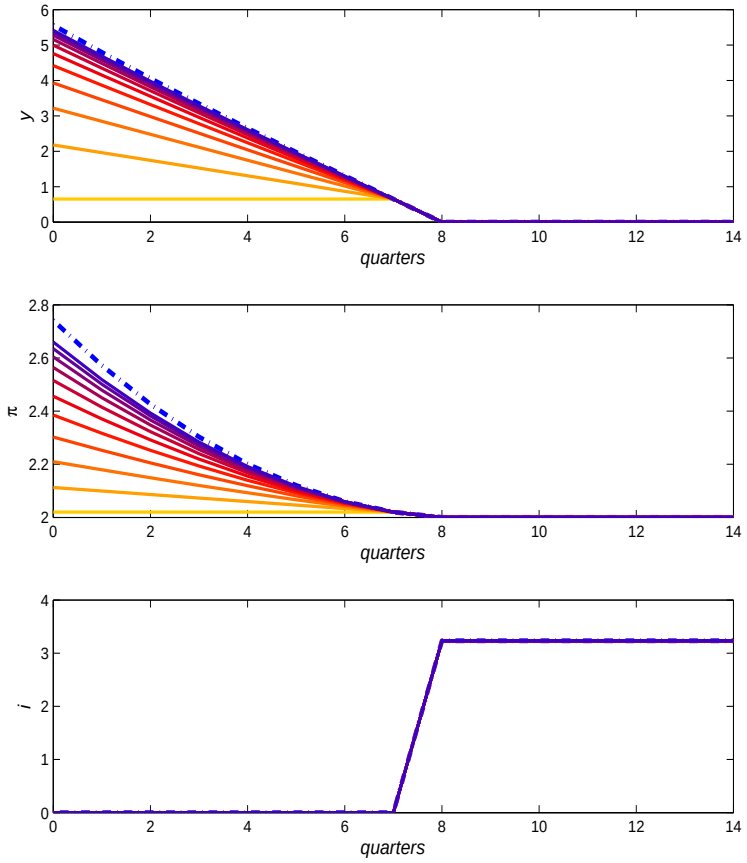

Figure 3: Reflective equilibrium outcomes for $n = 0$ through 4 compared with the FS-PFE solution, when the nominal interest rate is fixed for 8 quarters.

to remain fixed, because now there is no expectation of endogenous interest-rate increases in subsequent periods in response to the increases in output and inflation.

Because these stronger effects depend on reflection about what should happen in the future, given what is understood about future monetary policy, they are larger the greater the degree of reflection, and strongest under the assumption of perfect foresight. (They are also larger the longer the time for which the interest rate is expected to remain fixed, as this increases the degree to which reflection about the effects of future policy matters.) This means that in the case of a temporarily fixed interest rate, the difference between the PFE predictions and those obtained from a given finite degree of reflection is greater than that obtained in the case of a temporary shift in the Taylor-rule intercept.

In Figure 3, as in Figure 1, an average degree of reflection of $n = 4$ results in TE outcomes that are similar to the PFE predictions. But the reflective equilibrium outcomes when $n = 2$ are not as close to the PFE outcomes as they are in Figure 1, especially in the first quarters (when the anomalous policy is still expected to last for more than a year). In quarter zero, the output response when $n = 2$ is 14 percent smaller than the PFE prediction, and the inflation response is 10 percent smaller; and even when $n = 4$, the output and inflation responses are both about 3 percent smaller than the PFE predictions (whereas output differs by less than 1 percent in Figure 1, and inflation by less than 2 percent). Moreover, these discrepancies rapidly become much larger if the interest rate is expected to be fixed for an even longer period of time.

4.2 Very Long Periods with a Fixed Nominal Interest Rate

Much recent criticism of the implications of standard New Keynesian models regarding the effects of "forward guidance" have focused on the implications of such models (when solved under the assumption of perfect foresight or rational expectations) if one assumes that the nominal interest rate would be fixed for several years.[60] It should be noted that no central banks have actually experimented with date-based forward guidance that referred to dates more than about two years in the future; and while the period in which the U.S. federal funds rate target has remained at its lower bound has (as of the time of writing) lasted for more than six years, there was little reason for anyone to expect it to remain at this level for so long when the lower bound was reached at the end of 2008. Nonetheless, as discussed in section 1, thought experiments involving long-lasting periods at the ZLB remain useful for clarifying the theoretical coherence of proposed solution concepts.

If one assumes a date T many years in the future, the FS-PFE predicted effects on both output and inflation rapidly become extremely large. However, the effects predicted by reflective equilibrium with some modest (though positive) degree of reflection n do not grow in the same way, so that the PFE prediction rapidly becomes a worse and worse approximation to what one should expect in a reflective equilibrium with a modest level of n, if the horizon T is very long. Figure 4 illustrates this, in

[60] See, for example, Del Negro *et al.* (2013), Chung (2015), McKay *et al.* (2015), and Cochrane (2015a).

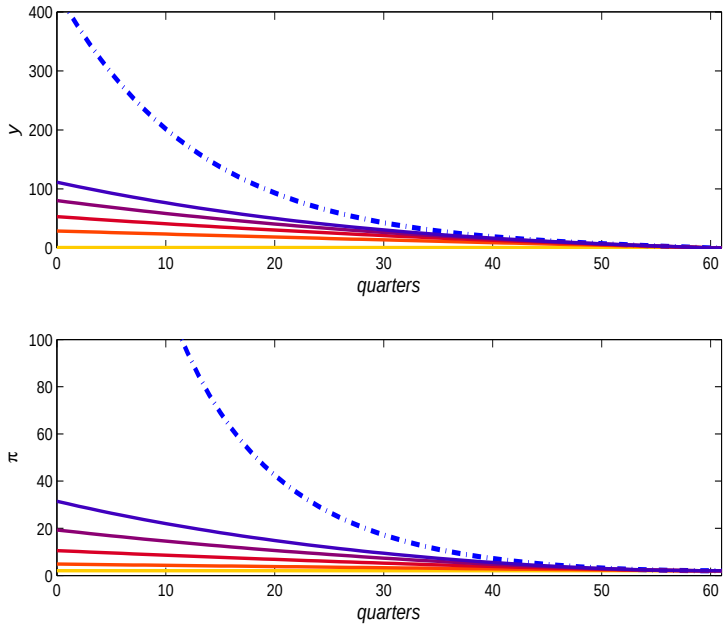

Figure 4: Reflective equilibrium outcomes for $n = 0$ through 4 compared with the FS-PFE solution, when the nominal interest rate is fixed for 15 years.

the case of the same model calibration as used in previous figures, by considering a (certainly unrealistic) situation in which the nominal interest rate is expected to be fixed for 15 years.[61]

According to the log-linearized model, an expectation of remaining at the ZLB for such a long time would, under the FS-PFE analysis, imply extremely large effects in the initial quarter: log output higher than its steady-state level by 4.36 (output 78 times its steady-state level), and an inflation rate of 442 percent.[62] Of course, such extreme predictions make it foolish to believe the assumptions made in this calculation (even given the assumption about policy): log-linearization of the model cannot be expected to yield even a roughly accurate result in the case of such a

[61]The third panel of the figure is omitted, since the expected path of the nominal interest rate is independent of the degree of reflection, as in Figure 3. Note also that now only the degrees of reflection $n = 0, 1, 2, 3, 4$ are shown in the figure, in order to allow the successive lines to be clearly distinguished from one another.

[62]Note that here, as in previous figures, π_t is reported as a conventional annualized rate, so that "$\pi_t = 100$" means that the price level will be twice as high (100 percent greater) after a year, while "$y_t = 100$" means that the log of output exceeds its steady-state value by 1.00, so that output is 2.72 times its steady-state level.

massive departure from steady-state conditions, nor do even the assumptions of the exact NK model — such as the assumption that the fraction $1 - \alpha$ of firms that do not reconsider their prices during the quarter simply supply whatever demand they receive at those prices — make sense under such extreme circumstances. We mention them only to point out that even granting the validity of our log-linearized model for purposes of such an exercise, the FS-PFE predictions are not at all a close approximation to the reflective equilibrium predictions.

Even if we assume $n = 4$ (a rather high average degree of reflection), the predicted increase in log output in quarter zero is instead only 1.11 (output 3 times its steady-state level), while the inflation rate is predicted to increase only to 31.5 percent per annum. If we assume a more modest degree of reflection, $n = 2$, the predicted increase in log output is only 0.53 (output 1.7 times its steady-state level), and inflation is predicted to increase only to 10.6 percent. This is still quite a large increase in output (large enough to make one doubt the realism of using the model for such an analysis), but these results are not close to the shocking predictions of the FS-PFE analysis.

The FS-PFE predictions of the log-linearized model become even more extreme if a longer period at the ZLB is contemplated: both the predicted effects on output and inflation grow without bound (and quite rapidly) as T is increased. For the kind of situation described in Proposition 4, but with $g_t = 0$ for all t, the FS-PFE paths for inflation and output are found by solving (2.13) for all $t < T$, working backward from the terminal condition $x_T = 0$ (which represents the unique FS-PFE given the specification for policy from date T onward). One obtains

$$x_\tau \;=\; -\sum_{j=0}^{\tau-1} B^j b\, \bar{\imath}_{SR} \tag{4.1}$$

for all $\tau \geq 1$, where $\tau \equiv T - t$ is again the number of periods remaining until policy is expected to revert to the Taylor rule, and the matrix B and vector b are the ones corresponding to policy response coefficients $\phi_\pi = \phi_y = 0$. We show in the Appendix that in this case the matrix B has an eigenvalue $\mu_2 > 1$, and that the left eigenvector e_2' associated with this eigenvalue satisfies $e_2' b \neq 0$. It follows that the solution (4.1) contains a component that grows as μ_2^τ as τ is made larger (which is to say, as T is made larger, for any value of t). Thus both elements of x_t grow exponentially as T is increased.[63]

[63]Note further that the elements of x_t are the *logarithm* of output and the *continuously com-*

Cochrane (2015a) objects to the FS-PFE as a solution concept on this ground, noting that it is implausible to suppose that changes in the specification of policy only very far in the future (say, a commitment to maintain the low interest rate for 1001 quarters instead of for only 1000 quarters) should have any significant effect on current economic outcomes. But this unpalatable feature of the FS-PFE is not a property of our concept of reflective equilibrium, assuming a fixed degree of reflection n as the length of the policy commitment is increased. Methods similar to those used to establish Proposition 2 also allow us to show the following.

Proposition 5 *Consider the case in which $g_t = 0$ for all t, and let the forward path of policy be specified as in Proposition 4. Then if the initial conjecture is given by $e_t(0) = 0$ for all t, the reflective equilibrium beliefs $\{e_t(n)\}$ for any degree of reflection n converge to a well-defined limiting value*

$$e_{SR}(n) \equiv \lim_{T \to \infty} e_t(n)$$

that is independent of n, and this limit is again given by (3.10), where $\bar{e}_{SR}^{PF}$ is again defined in (3.11). The reflective equilibrium outcomes for output, inflation and the nominal interest rate then converge as well as T is made large, to the values obtained by substituting the beliefs $e_{SR}(n)$ into the TE relations (2.10) and the reaction function (2.8).

The proof is given in the Appendix. The result is similar to the one stated in Proposition 2.[64] There is one important difference, however: in the present case, the stationary expectations $\bar{e}_{SR}^{PF}$ no longer correspond to a unique FS-PFE associated with permanent maintenance of the interest rate $i_t = \bar{\imath}_{SR}$. (There is no unique FS-PFE under such a policy; instead, as discussed in section 2.2, there is a continuum of PFE that all converge asymptotically to the steady state in which expectations are given by $\bar{e}_{SR}^{PF}$.)

pounded rate of inflation; these quantities must both be exponentiated to obtain the level of output and the factor by which prices increase relative to the previous year's prices. Thus even if the elements of x_t only grew *linearly* with T, output and the conventional measure of inflation would both grow exponentially. Instead, here the latter quantities grow as the exponential of an exponential.

[64] Note that Proposition 2, as stated earlier, did not require that the reaction function coefficients satisfy (2.14); it would apply, in particular, to the case $\phi_\pi = \phi_y = 0$, corresponding to fixed interest rates before and after date T. The only difference here is that Proposition 5 establishes a similar result even when the response coefficients prior to date T differ from those from date T onward.

Thus if we consider the reflective equilibrium associated with any given finite degree of reflection n, we find that equilibrium outcomes are essentially the same for any long enough horizon T. Moreover, for any long enough T, reflective equilibrium outcomes are nearly constant over time, and close to the constant outcomes that occur under a reflective equilibrium of the same degree in the case of a permanent commitment to the fixed interest rate $i_t = \bar{\imath}_{SR}$. (This last observation follows from a comparison of (3.10)–(3.11) with (3.6)–(3.7), where the latter equations define the reflective equilibrium of degree n in the case of a permanent commitment to a given reaction function.) Thus there is no material difference, as far as reflective equilibrium is concerned, between commitment to a fixed interest rate for a long but finite time and a commitment to fix the interest rate permanently.

This attractive feature of reflective equilibrium does not, however, mean that it leads to predictions similar to those of Cochrane's (2015a) "backward stable" PFE solution. In those cases where the degree of reflection n is large enough for the reflective equilibrium to correspond nearly to a PFE (as, for example, in the case that $n = 4$ or larger, for the parameter values and policy experiment considered in Figure 3), the PFE that it approximates is the FS-PFE (uniquely defined in the case of a finite-length interest-rate peg), and *not* the backward-stable PFE. These solutions are in fact quite different — not only in the case of large values of T, but even when T is very short.[65] It is thus important to note that one need not accept Cochrane's solution concept as a sensible one, in order to avoid the unpalatable prediction of explosive behavior as T is made large.

[65]They imply quite different equilibrium responses even when T is *arbitrarily* short: in a continuous-time version of the model, they would imply different responses even in the limit as the continuous length of time T is made infinitesimally small (which is possible because under the "backward stable" solution, outcomes *after* date T depend on the policy pursued *before* that date). This is one of the especially unattractive features of the "backward stable solution" as a solution concept: it implies that pegging the interest rate at different levels should lead to different equilibrium outcomes over a period of years, even when the pegs in question are to last for only one second! The concept of a reflective equilibrium for some given degree of reflection n avoids this undesirable prediction, while *also* yielding predictions that converge as T is made unboundedly large.

4.3 The Paradox Explained

We can now explain the error in the reasoning sketched in the introduction. It is true that under the assumption of a *permanent* interest-rate peg, the only forward-stable PFE are ones that converge asymptotically to an inflation rate determined by the Fisher equation and the interest-rate target (and thus, lower by one percentage point for every one percent reduction in the interest rate). But for most possible initial conjectures (as starting points for the process of belief revision proposed above), *none* of these perfect foresight equilibria correspond, even approximately, to reflective equilibria — even to reflective equilibria for some very high degree of reflection n. Nor is this because in such cases high-n reflective equilibria correspond to some *other* kind of PFE; instead, one generally finds that the belief-revision dynamics *fail to converge* to *any* PFE as n increases, in the case of a permanent interest-rate peg.

This failure of convergence can be illustrated using results already presented above. In the case of a policy under which $i_t = \bar{\imath}_{LR}$ forever, if we further assume that $g_t = 0$ for all t and start from an initial conjecture under which $e_t = 0$ for all t, then the belief-revision dynamics are given by (3.5) for all t, where M in this equation is now the matrix corresponding to response coefficients $\phi_\pi = \phi_y = 0$, and we now have $e_t(n) = e_{LR}(n)$ for all t.[66] The solution for general n is again given by (3.6), where $\bar{e}_{LR}^{PF}$ is again defined by (3.7). However, whereas in the Taylor-rule case considered in section 3, this solution implied that $e_{LR}(n) \to \bar{e}_{LR}^{PF}$ as $n \to \infty$, this is no longer true in the case of an interest-rate peg. When $\phi_\pi = \phi_y = 0$, we show in the Appendix that the matrix $M - I$ has a positive real eigenvalue. This in turn means that the elements of the matrix $\exp[n(M - I)]$ grow explosively as n is made large, and $e_{LR}(n)$ diverges from $\bar{e}_{LR}^{PF}$, rather than converging to it. Nor does $e_{LR}(n)$ approach any PFE: the distance between $e_{LR}(n)$ and $e_{LR}^*(n)$ also grows explosively as n increases.

It similarly follows (using Proposition 5) that the nearly-stationary outcomes obtained in the case of any long enough finite-length interest-rate peg under a fixed degree of reflection n do not converge to any limit as n is made large. Thus neither of the double limits

$$\lim_{n\to\infty} \lim_{T\to\infty} e_t(n) \;=\; \lim_{n\to\infty} e_{SR}(n)$$

[66] The case considered now is of the same kind as was considered in deriving (3.5), except that we now set $T = 0$, and assume $\phi_\pi = \phi_y = 0$.

or

$$\lim_{T\to\infty}\lim_{n\to\infty} e_t(n) \;=\; \lim_{T\to\infty} e_t^{PF}$$

is well-defined in the case of a temporary interest-rate peg.[67] It is true (for any finite length of peg) that a high enough degree of reflection leads to an outcome indistinguishable from a forward-stable PFE; and it is also true (for any finite degree of reflection) that a long enough finite-length peg leads to reflective equilibrium outcomes that are indistinguishable from those under a permanent peg. But it does *not* follow from these observations that a long enough peg together with a high enough degree of reflection must lead to anything similar to a forward-stable PFE associated with a permanent interest-rate peg. It is the failure to recognize this that leads to paradoxical conclusions in the argument sketched in the introduction.

4.4 Consequences of Maintaining a Low Interest Rate for Longer

Consideration of the possible PFE in the case of a permanently fixed interest rate thus need not provides a correct conclusion as to the likely effects of a commitment to maintain the nominal interest rate at a low level for a longer time than that for which the zero lower bound prevents a central bank from implementing its normal reaction function. In fact, one can easily show that for any given degree of reflection, commitment to keep the nominal interest rate at a low level for a longer period of time is necessarily both expansionary and inflationary, at least in the case where (at the future horizon at which one is lengthening the commitment to fixed-interest-rate policy) neither exogenous disturbances nor the assumed initial conjecture are sources of deflationary pressures.

Proposition 6 *For a given shock sequence* $\{g_t\}$ *and a given initial conjecture* $\{e_t(0)\}$*, consider monetary policies of the kind described in Proposition 4, with* $\bar{\imath}_{SR} < 0$ *(that is, an initial fixed interest rate at a level lower than the steady-state nominal interest rate associated with the long-run inflation target* π^**). Suppose also that* $g_t = 0$ *and*

[67] Note that $\bar{e}_{SR}^{PF}$, the common limit given in Proposition 3, is still well-defined in this case. But $e_{SR}(n)$ no longer converges to it as n is made large, nor does e_t^{PF} converge to it as T is made large. Failure of the "Taylor Principle" invalidates *both* of those convergence results, relied upon in Proposition 3.

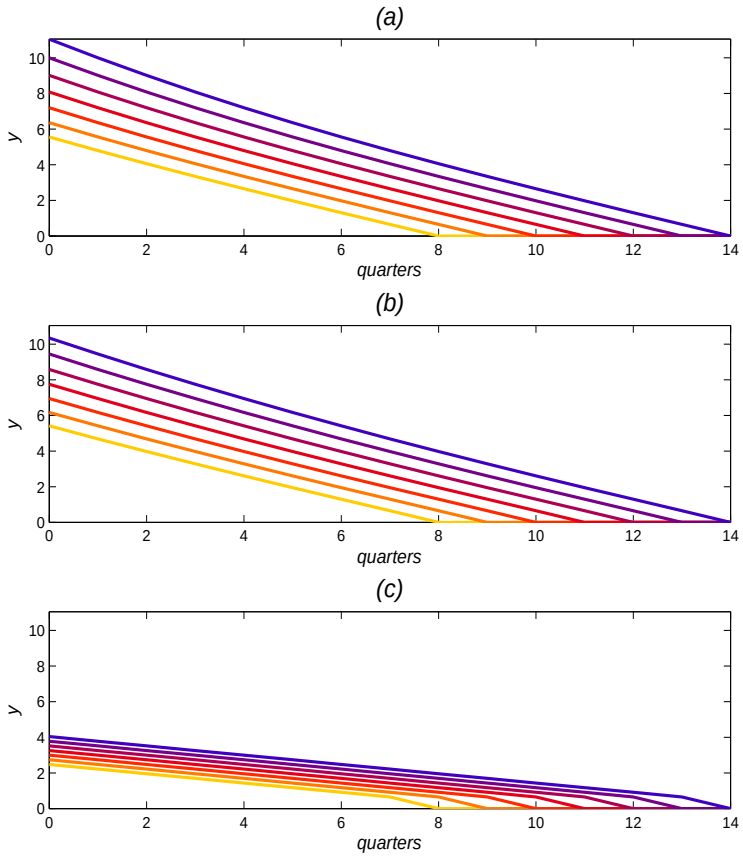

Figure 5: Output responses if the interest rate is fixed at zero for T quarters, where T takes values between 8 and 14. Panel (a): PFE; panel (b): $n = 4$; panel (c): $n = 0.5$.

$e_t(0) = 0$ *for all* $t \geq T$.[68] *Then for any fixed* $\bar{\imath}_{SR}$ *and fixed level of reflection* $n > 0$, *increasing the length of the commitment from* T *to* $T' > T$ *increases both inflation and output in the reflective equilibrium, in all periods* $0 \leq t < T'$, *while it has no effect on either variable from date* T' *onward.*

The proof is given in the Appendix.

Figure 5 illustrates the effect indicated in Proposition 6, for the case of a pure shift in monetary policy (that is, one in which $g_t = 0$ and $e_t(0) = 0$ for all t).

[68] In fact, it should be evident from the proof given in the Appendix that it suffices that $g_t \geq 0$ and $e_t(0) \geq 0$ for all $t \geq T$. What matters for the proof is that there not be factors tending to reduce output or inflation, apart from the effects of monetary policy, that are anticipated to affect periods beyond date T.

Model parameters are as in the earlier numerical examples, and as in Figures 3 and 4, $\bar{\imath}_{SR}$ is set at the lowest rate allowed by the zero lower bound. In each panel, the equilibrium paths for output are compared in the case of alternative values of T, ranging from 8 quarters up to 14 quarters. The successive panels indicate the outcomes under different degrees of reflection: in panel (a), the FS-PFE outcomes are shown (corresponding to the limit as $n \to \infty$, given Proposition 4); in panel (b), the reflective equilibrium outcomes for the case $n = 4$; and in panel (c), the corresponding outcomes if instead $n = 0.5$.

One sees that with each successive increase in the length of time for which the low interest rate is to be maintained, output is increased, in each of the periods in which the interest rate is fixed; this is true regardless of the assumed value of n.[69] (Inflation is similarly increased, though we do not show the corresponding responses for inflation.) In the case of a high enough degree of reflection (such as the case $n = 4$, shown in the figure), the reflective equilibrium outcomes are similar to the FS-PFE outcomes. But even when the degree of reflection is much lower, the outcomes *qualitatively* resemble those predicted by the FS-PFE analysis, even if the quantitative magnitude of the effects is quite different.

The *quantitative* effects can, however, be quite different from those implied by the FS-PFE analysis; they are particularly different in the case of long periods with a fixed interest rate. Indeed, while the FS-PFE analysis implies that the effects of *any* contractionary shock, no matter how severe, can necessarily be completely counteracted by a sufficiently long-lasting commitment to a low interest rate (albeit one that remains non-negative) — and in fact, that a sufficiently long-lasting commitment can produce an inflationary boom of arbitrary size — it is possible, under the reflective equilibrium analysis, to find (if the degree of reflection is small enough) that even a promise to keep the interest rate *permanently* at zero would be insufficient to prevent output and inflation from both falling below their target values. Proposition 5 implies that there will be only a finite amount of stimulus provided even by a permanent interest-rate peg, and this need not be enough to prevent output and inflation from falling in response to a disturbance.

[69] Except, of course, in the limiting case $n = 0$. When $n = 0$, as illustrated in Figure 3, the effects on output and inflation are independent of the number of remaining periods for which the interest rate is expected to be fixed, as expectations regarding future policy have no influence.

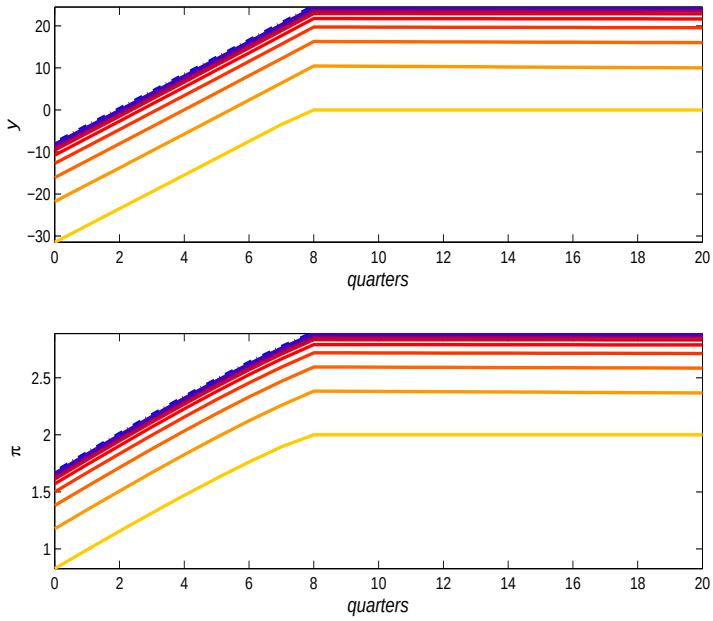

Figure 6: Output and inflation responses in a reflective equilibrium with $n = 0.1$, when ρ_t is reduced for 8 quarters, and the interest rate is kept at zero for a period of length T. Darker lines indicate progressively longer periods T (to infinity).

Figure 6 illustrates this possibility. The model parameters are the same as in the previous numerical illustrations, but we now consider a real disturbance that lowers the discount rate ρ_t by 5 percentage points per quarter, and that lasts for 8 quarters. (The discount rate returns to its low normal value again in quarter 8.) This is a "Great Depression" magnitude of disturbance: as shown in the figure, in the absence of any commitment to depart from the normal reaction function after quarter 8 (the time at which it becomes possible again to implement a standard Taylor rule), this disturbance causes output to contract by more than 30 percent. The figure shows the responses of output and inflation to this shock, under a variety of assumptions about the length of time T for which it is announced that the nominal interest rate will be held at its lower bound, after which the central bank will revert to its normal (Taylor-rule) reaction function. In each case, the outcomes shown are for a reflective equilibrium in which the degree of reflection is only $n = 0.1$.[70]

[70]This is quite a low level of reflection, but is chosen to illustrate our point.

The lightest lines correspond to the case $T = 8$ quarters: that is, the interest rate will remain at the lower bound only for as long as the Taylor rule would require an even lower rate than that (which cannot be implemented). Progressively darker lines indicate the effects of progressively longer commitments; the lines shown are for periods of 50 years, 100 years, 150 years and so on. The final dash-dotted line indicates the effect of a commitment to keep the interest rate permanently at the zero lower bound. In accordance with Proposition 6, each lengthening of the commitment increases both output and inflation; but, in accordance with Proposition 5, the outcomes associated with all long enough commitments converge to the outcomes predicted in the case of a permanent commitment. In this example (involving a very low, though positive, degree of reflection), even the permanent commitment is insufficient to prevent both output and inflation from falling below their target values in the quarter of the shock (quarter zero), though the long-lasting low-interest-rate regimes result in very substantial output booms, and persistently above-target inflation, later on (that last for decades).

Thus while our reflective equilibrium analysis confirms the result of PFE analyses using the conventional (FS-PFE) equilibrium selection, according to which a commitment to keep the interest rate at its lower bound for a longer time should be expansionary, it also indicates that — given that it is realistic to assume that people would truncate the belief-revision process at some finite level of reflection, and quite possibly at a relatively low one — rational-expectations analyses almost certainly overstate the magnitude of stimulus that one can expect to obtain from such commitments, even when understood and believed by all individuals. This can be added to the varied list of reasons that other authors have proposed for doubting that forward guidance should be as extraordinarily powerful as rational-expectations analyses using highly forward-looking NK models sometimes suggest.[71] While our analysis still implies that commitments of this kind should provide a potentially powerful tool, of particular usefulness when a central bank is constrained by the zero lower bound, it increases the possible scope for using other tools, such as fiscal policy, under such circumstances as well.

[71] Again see Del Negro *et al.* (2013), Chung (2015), and McKay *et al.* (2015).

5 Conclusion

Is there, then, reason to fear that a commitment to keep nominal interest rates low for a longer period of time will be deflationary, rather than inflationary? There is one way in which such an outcome could easily occur, and that is if the announcement of the policy change were taken to reveal negative information (previously known only to the central bank) about the outlook for economic fundamentals, rather than representing a pure change in policy intentions of the kind analyzed above.[72] This may well have been a problem with the way in which "date-based forward guidance" was used by the U.S. Federal Reserve during the period 2011-12, as discussed by Woodford (2012); but it is not an inherent problem with announcing a change in future policy intentions, only with a particular way of explaining what has changed.

The idea that a commitment to keep nominal interest rates low for a longer time should be deflationary, even when understood to represent a pure change in monetary policy — simply because the only rational-expectations equilibria in which nominal interest rates remain forever low involve deflation — is instead mistaken, in our view. If people believe the central bank's statements about its future policy intentions, and believe that it will indeed succeed in maintaining a low nominal interest rate, it does not follow that they must expect a deflationary equilibrium; this does not follow, even if we suppose that they reason about the economy's likely future path using a correct model of how inflation and aggregate output are determined.

If their reasoning occurs through a process of reflection of the kind modeled in this paper, then an increase in the expected length of time for which the nominal interest rate is expected to remain at some effective lower bound should result in expectations of higher income and higher inflation, regardless of the degree of reflection (as long as $n > 0$); and according to our model of temporary equilibrium resulting from optimizing spending and pricing decisions, such a change in expectations should result in higher output and inflation. This outcome may or may not approximate the outcome associated with a perfect foresight equilibrium, depending on the degree of reflection; in the case of a commitment to keep the nominal interest rate low for a long enough period, it almost certainly will *not* resemble any PFE, even approximately.

This is why it is important to explicitly model the process of belief revision as a

[72] For further discussion of the way in which the revelation of central-bank information by announced policy decisions can result in perverse effects, see García-Schmidt (2015).

result of further reflection, rather than simply assuming that the PFE must yield a correct prediction. Some macroeconomists may find the proposed alternative solution concept (reflective equilibrium for some finite degree of reflection n) unappealing, on the ground that it yields a less definite prediction than the assumption of perfect foresight (or rational expectations) equilibrium. But while it is true that our conclusions about the effects of a given policy commitment depend both on the exact choice of an initial conjecture and on exactly how far one supposes that people should continue the belief-revision process, this does not mean that we are unable to draw any conclusions of relevance to policy deliberations. Our conclusions as to the *signs* of the effects just mentioned are independent of those details of the specification of the reflective equilibrium. Hence it is possible to obtain conclusions of a useful degree of specificity even when one has little ground for insisting on a single precise model of expectation formation.

It should also be noted that while our concept of reflective equilibrium can yield quite various predictions (for differing assumptions about the initial conjecture and the degree of reflection) under some circumstances, because the belief-revision dynamics diverge (or converge quite slowly), under other circumstances much tighter predictions are obtained, because of relatively rapid convergence of the belief-revision dynamics. It can then be a goal of policy design to choose a policy with the property that the belief-revision dynamics should converge reliably, leading to less uncertainty about the outcome that should be expected under the policy.

In the case of a central bank that finds itself seeking additional demand stimulus when it has already cut its short-term nominal interest rate instrument to its effective lower bound, a commitment to maintain the instrument at the lower bound for a long time *that can be announced in advance,* regardless of how economic conditions develop, is *not* an ideal policy response, according to this criterion. Such a policy should be expected to be stimulative, according to the analysis in this paper; but the exact degree of stimulus is difficult to predict. It may not be possible to choose a length of time for which to commit to the ultra-low interest rate that does not run simultaneously the risk of being too short to be effective, if the degree of reflection n is too low, and the opposite risk of being wildly inflationary, if the degree of reflection n is too high.

But one could achieve a less uncertain outcome, according to the reflective equilibrium analysis, by committing to maintain a low nominal interest rate until some

macroeconomic target is reached, such as the price-level target proposed by Eggertsson and Woodford (2003), or the nominal-GDP target path proposed by Woodford (2012).[73] In the case that people carry the belief-revision process forward to a high degree, they should expect interest rates to be raised relatively soon, under such a commitment; but if instead they truncate the process at a relatively low degree of reflection, they should expected interest rates to remain low for much longer. In either case, belief that the central bank is serious about the policy should change expectations in a way that results in a substantial, but not extravagant, increase in current aggregate demand.

Thus even though the approach proposed here leads to a *set* of possible predictions in the case of a given policy specification rather than a *point* prediction, this does not mean that the approach yields no conclusions that are useful for policy design. Instead, insisting on the use of perfect foresight equilibrium analysis simply because it yields a more precise prediction may lead to large errors. One is reminded of the dictum of the British logician Carveth Read:[74] "It is better to be vaguely right than exactly wrong."

[73]This alternative to date-based forward guidance would also have the advantage of being less likely to be misunderstood as revealing negative central-bank information about fundamentals, as discussed by Woodford (2012).

[74]Read (1920), p. 351. The aphorism is often mis-attributed to John Maynard Keynes.

References

[1] Arad, Ayala, and Ariel Rubinstein, "The 11-20 Money Request Game: A Level-k Reasoning Study," *American Economic Review* 102: 3561-3573 (2012).

[2] Bullard, James, "Seven Faces of 'The Peril'," *Federal Reserve Bank of St. Louis Review* 92(5): 339-352 (2010).

[3] Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong, "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics* 119: 861-898 (2004).

[4] Chung, Hess, "The Effects of Forward Guidance in Three Macro Models," *FEDS Notes,* Federal Reserve Board, February 26, 2015. [URL: http://www.federalreserve.gov/econresdata/notes/feds-notes/2015/effects-of-forward-guidance-in-three-macro-models-20150226.html]

[5] Cochrane, John H., "Determinacy and Identification with Taylor Rules," *Journal of Political Economy* 119: 565-615 (2011).

[6] Cochrane, John H., "Monetary Policy with Interest on Reserves," unpublished, October 2014.

[7] Cochrane, John H., "The New-Keynesian Liquidity Trap," unpublished, revised January 2015a.

[8] Cochrane, John, "Doctrines Overturned," *The Grumpy Economist,* February 28, 2015b. [URL: http://johnhcochrane.blogspot.com/2015/02/doctrines-overturned.html]

[9] Crawford, Vincent P., Miguel A. Costa-Gomes, and Nagore Iriberri, "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence and Applications," *Journal of Economic Literature* 51: 5-62 (2013).

[10] Del Negro, Marco, Marc Giannoni, and Christina Patterson, "The Forward Guidance Puzzle," Federal Reserve Bank of New York Staff Report no. 574, revised May 2013.

[11] Denes, Matthew, Gauti B. Eggertsson, and Sophia Gilbukh, "Deficits, Public Debt Dynamics, and Tax and Spending Multipliers," *Economic Journal* 123: F133-F163 (2013).

[12] Eggertsson, Gauti B., and Michael Woodford, "The Zero Bound on Interest Rates and Optimal Monetary Policy," *Brookings Papers on Economic Activity* 2003(1): 139-211.

[13] Evans, George W., and Seppo Honkapohja, *Learning and Expectations in Macroeconomics,* Princeton: Princeton University Press, 2001.

[14] Evans, George W., and Garey Ramey, "Expectation Calculation and Macroeconomic Dynamics," *American Econ. Review* 82: 207-224 (1992).

[15] Evans, George W., and Garey Ramey, "Expectation Calculation, Hyperinflation and Currency Collapse," in H. Dixon and N. Rankin, eds., *The New Macroeconomics: Imperfect Markets and Policy Effectiveness,* Cambridge: Cambridge University Press, 1995.

[16] Evans, George W., and Garey Ramey, "Calculation, Adaptation and Rational Expectations," *Macroeconomic Dynamics* 2: 156-182 (1998).

[17] Fair, Ray, and John B. Taylor, "Solution and Maximum Likelihood Estimation of Dynamic Nonlinear Rational Expectations Models," *Econometrica* 51: 1169-1185 (1983).

[18] Friedman, Milton, "The Role of Monetary Policy," *American Economic Review* 58: 1-17 (1968).

[19] García-Schmidt, Mariana, "Monetary Policy Surprises and Expectations," unpublished, Columbia University, August 2015.

[20] Guesnerie, Roger, "An Exploration of the Eductive Justifications of the Rational Expectations Hypothesis," *American Economic Review* 82: 1254-1278 (1992).

[21] Guesnerie, Roger, "Macroeconomic and Monetary Policies from the Eductive Viewpoint," in K. Schmidt-Hebbel and C. Walsh, eds., *Monetary Policy Under Uncertainty and Learning,* Santiago: Central Bank of Chile, 2008.

[22] Hirsch, Morris W., and Stephen Smale, *Differential Equations, Dynamical Systems, and Linear Algebra,* New York: Academic Press, 1974.

[23] Jury, Eliahu I., *Theory and Application of the z-Transform Method,* New York: Wiley, 1964.

[24] Keynes, John Maynard, *The General Theory of Employment, Interest and Money,* London: Macmillan, 1936.

[25] McCallum, Bennett T., "On Non-Uniqueness in Rational Expectations Models: An Attempt at Perspective," *Journal of Monetary Economics* 11: 139-168 (1983).

[26] McCallum, Bennett T., "Role of the Minimal State Variable Criterion in Rational Expectations Models," *International Tax and Public Finance* 6: 621-639 (1999).

[27] McKay, Alisdair, Emi Nakamura, and Jon Steinsson, "The Power of Forward Guidance Revisited," unpublished, July 2015.

[28] Nagel, Rosemarie, "Unraveling in Guessing Games: An Experimental Study," *American Economic Review* 85: 1313-1326 (1995).

[29] Phelps, Edmund S., "The Trouble with 'Rational Expectations' and the Problem of Inflation Stabilization," in R. Frydman and E.S. Phelps, eds., *Individual Forecasting and Aggregate Outcomes,* Cambridge: Cambridge University Press, 1983.

[30] Read, Carveth, *Logic: Deductive and Inductive,* London: Simkin, Marshall, 1920.

[31] Schmitt-Grohé, Stephanie, and Martín Uribe, "Liquidity Traps: An Interest-Rate-Based Exit Strategy," NBER Working Paper no. 16514, November 2010.

[32] Stahl, Dale O., "Boundedly Rational Rule Learning in a Guessing Game," *Games and Economic Behavior* 16: 303-330 (1996).

[33] Stahl, Dale O., and Paul W. Wilson, "Experimental Evidence on Players' Models of Other Players," *Journal of Economic Behavior and Organization* 25: 309-327 (1994).

[34] Stahl, Dale O., and Paul W. Wilson, "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior* 10: 218-254 (1995).

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

[36] Werning, Ivan, "Managing a Liquidity Trap: Monetary and Fiscal Policy," unpublished, April 2012.

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

[38] Woodford, Michael, "Methods of Policy Accommodation at the Interest-Rate Lower Bound," in *The Changing Policy Landscape,* Kansas City: Federal Reserve Bank of Kansas City, 2012.

[39] Woodford, Michael, "Macroeconomic Analysis Without the Rational Expectations Hypothesis," *Annual Review of Economics* 5: 303-346 (2013).

[40] Yun, Tack, "Nominal Price Rigidity, Money Supply Endogeneity, and Business Cycles," *Journal of Monetary Economics* 37: 345-370 (1996).

APPENDIX

A Matrices of Coefficients and their Properties

A.1 Temporary Equilibrium Solution

The system of three equations given in the text can be solved to obtain

$$x_t = Ce_t + c\omega_t \tag{A.1}$$

where we define the vectors

$$x_t = \begin{bmatrix} y_t \\ \pi_t \end{bmatrix}, \qquad e_t = \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}, \qquad \omega_t = \begin{bmatrix} g_t \\ \bar{\imath}_t \end{bmatrix},$$

and the matrices

$$C = \tfrac{1}{\Delta} \begin{bmatrix} 1 & -\sigma\phi_\pi(1-\alpha)\beta \\ \kappa & (1+\sigma\phi_y)(1-\alpha)\beta \end{bmatrix}, \qquad c = \tfrac{1}{\Delta} \begin{bmatrix} 1 & -\sigma \\ \kappa & -\kappa\sigma \end{bmatrix},$$

and use the shorthand notation $\Delta \equiv 1 + \sigma\phi_y + \sigma\kappa\phi_\pi \geq 1$. (This last inequality, that allows us to divide by Δ, holds under the sign restrictions maintained in the text.) Given this solution for x_t, the solution for the nominal interest rate is obtained by substituting the solutions for inflation and output into the reaction function (2.8).

This solution also allows us to solve for the summary variables a_t that decision-makers need to forecast, resulting in

$$a_t = Me_t + m\omega_t$$

where we define

$$M = \tfrac{1}{\Delta} \begin{bmatrix} \frac{1+\sigma\kappa-\beta\Delta}{1-\beta} & \frac{\sigma\beta(1-\alpha)(1+\sigma\phi_y-\phi_\pi)}{1-\beta} \\ \frac{\kappa}{(1-\alpha)(1-\alpha\beta)} & \frac{\beta(1+\sigma\phi_y-\alpha\Delta)}{1-\alpha\beta} \end{bmatrix}, \qquad m = \tfrac{1}{\Delta} \begin{bmatrix} \frac{1+\sigma\kappa-\beta\Delta}{1-\beta} & -\frac{\sigma(1+\sigma\kappa)}{1-\beta} \\ \frac{\kappa}{(1-\alpha)(1-\alpha\beta)} & -\frac{\sigma\kappa}{(1-\alpha)(1-\alpha\beta)} \end{bmatrix}.$$

A.2 Perfect Foresight Equilibrium Dynamics

It follows from the discussion in the text (citing Woodford, 2003, chap. 4) that the PFE dynamics can be written in the form

$$x_t \;=\; B\,x_{t+1} + b\,(\rho_t - \bar{\imath}_t) \tag{A.2}$$

where we define

$$B = \tfrac{1}{\Delta} \begin{bmatrix} 1 & \sigma(1-\beta\phi_\pi) \\ \kappa & \sigma\kappa + \beta(1+\sigma\phi_y) \end{bmatrix}, \qquad b = \tfrac{1}{\Delta} \begin{bmatrix} \sigma \\ \sigma\kappa \end{bmatrix}.$$

Alternatively, we can characterize PFE dynamics by the requirement that e_t must equal e_t^* for all t. From (2.20) it follows that a sequence of vectors of expectations $\{e_t\}$ constitute PFE expectations if and only if

$$\begin{aligned} e_t &= e_t^* = \sum_{j=1}^{\infty} \psi_j e_{t+j} + \sum_{j=1}^{\infty} \varphi_j \omega_{t+j} \\ &= \psi_1 e_{t+1} + \varphi_1 \omega_{t+1} + \Lambda e_{t+1} \\ &= (I-\Lambda)M e_{t+1} + (I-\Lambda)m\,\omega_{t+1} + \Lambda e_{t+1} \\ &= [(I-\Lambda)M + \Lambda] e_{t+1} + (I-\Lambda)m\,\omega_{t+1} \end{aligned} \tag{A.3}$$

for all $t \geq 0$.

The dynamics implied by (A.3) are in fact equivalent to those implied by (A.2). Using (A.1) together with (A.3) implies that the PFE dynamics of output and inflation must satisfy

$$\begin{aligned} x_t &= C[(I-\Lambda)M + \Lambda] e_{t+1} + C(I-\Lambda)m\,\omega_{t+1} + c\,\omega_t \\ &= C[(I-\Lambda)M + \Lambda] C^{-1} [x_{t+1} - c\omega_{t+1}] + C(I-\Lambda)m\,\omega_{t+1} + c\,\omega_t. \end{aligned}$$

But this relation is in fact equivalent to (A.2), given that our definitions above imply that

$$\begin{aligned} C[(I-\Lambda)M + \Lambda] C^{-1} &= B, \\ C(I-\Lambda)m &= Bc + b \cdot [-\beta\sigma^{-1} \;\; 0], \\ c &= b \cdot [\sigma^{-1} \;\; -1]. \end{aligned} \tag{A.4}$$

A.3 Properties of the Matrix M

A number of results turn on the eigenvalues of the matrix

$$M - I = \frac{1}{\Delta} \begin{bmatrix} -\frac{\sigma\phi_y + \sigma\kappa\phi_\pi - \sigma\kappa}{1-\beta} & \frac{(1-\alpha)\sigma\beta(1+\sigma\phi_y-\phi_\pi)}{1-\beta} \\ \frac{\kappa}{(1-\alpha)(1-\alpha\beta)} & \frac{\beta(1+\sigma\phi_y)-\Delta}{1-\alpha\beta} \end{bmatrix}.$$

We first note that the determinant of the matrix is given by

$$\mathrm{Det}(M-I) = \frac{\sigma\kappa}{\Delta(1-\beta)(1-\alpha\beta)} \left(\phi_\pi + \frac{(1-\beta)}{\kappa}\phi_y - 1 \right).$$

Under our sign assumptions, the factor pre-multiplying the factor in parentheses is necessarily positive. Hence the determinant is non-zero (and the matrix is non-singular) if

$$\phi_\pi + \frac{(1-\beta)}{\kappa}\phi_y - 1 \neq 0. \tag{A.5}$$

(In this case the steady-state vector of expectations (3.7) is well-defined, as asserted in the text.)

For any 2×2 real matrix A, both eigenvalues have negative real part if and only if $\text{Det}[A] > 0$ and $\text{Tr}[A] < 0$.[75] From the result above, the first of these conditions is satisfied if the left-hand side of (A.5) is positive, which is to say, if the Taylor Principle (2.14) is satisfied. The trace of $M - I$ is given by

$$\text{Tr}(M - I) \;=\; -\frac{1}{\Delta}\left(\frac{\sigma(\phi_y + \kappa\phi_\pi - \kappa)}{1-\beta} + \frac{\sigma\kappa\phi_\pi + (1-\beta)(1+\sigma\phi_y)}{1-\alpha\beta}\right).$$

The second term inside the parentheses is necessarily positive under our sign assumptions, and the first term is positive as well if the Taylor Principle is satisfied, since

$$\phi_y + \kappa\phi_\pi - \kappa \;=\; \kappa\left(\phi_\pi + \frac{\phi_y}{\kappa} - 1\right) \;>\; \kappa\left(\phi_\pi + \frac{\phi_y(1-\beta)}{\kappa} - 1\right) \;>\; 0. \qquad \text{(A.6)}$$

Hence the Taylor Principle is a sufficient condition for $\text{Tr}[M - I] < 0$. It follows that (given our other sign assumptions) the Taylor Principle is both necessary and sufficient for both eigenvalues of $M - I$ to have negative real part.

If instead the left-hand side of (A.5) is negative, $\text{Det}[M - I] < 0$, and as a consequence the matrix must have two real eigenvalues of opposite sign.[76] Thus one eigenvalue is positive in this case, as asserted in the text. Note that this is the case that obtains if $\phi_\pi = \phi_y = 0$.

A.4 A Further Implication of the Taylor Principle

We are also interested in the eigenvalues of the related matrix $A(\lambda)M - I$, where for an arbitrary real number $-1 \leq \lambda \leq 1$, we define

$$A(\lambda) \;\equiv\; \begin{pmatrix} \frac{\lambda(1-\delta_1)}{1-\lambda\delta_1} & 0 \\ 0 & \frac{\lambda(1-\delta_2)}{1-\lambda\delta_2} \end{pmatrix}.$$

(Note that in the limiting case $\lambda = 1$, this reduces to the matrix $M - I$, just discussed.) In the case that the Taylor principle (2.14) is satisfied, we can show that for any $-1 \leq \lambda \leq 1$, both eigenvalues of $A(\lambda)M - I$ have negative real part. This follows again from a consideration of the determinant and trace of the matrix (generalizing the above discussion).

[75] See, for example, Hirsch and Smale (1974), p. 96.

[76] Again see Hirsch and Smale (1974), p. 96.

Since

$$A(\lambda)M - I \;=\; \frac{1}{\Delta}\begin{bmatrix} -\frac{\Delta-\lambda(1+\sigma\kappa)}{1-\beta\lambda} & -\frac{\sigma(1-\alpha)\beta(\phi_\pi-1-\sigma\phi_y)\lambda}{1-\beta\lambda} \\ \frac{\kappa\lambda}{(1-\alpha)(1-\alpha\beta\lambda)} & -\frac{\Delta-\beta\lambda(1+\sigma\phi_y)}{1-\alpha\beta\lambda} \end{bmatrix},$$

we have

$$\mathrm{Det}(A(\lambda)M - I) \;=\; \frac{\Delta - \lambda(\beta(1+\sigma\phi_y)+1+\sigma\kappa)+\beta\lambda^2}{\Delta(1-\beta\lambda)(1-\alpha\beta\lambda)}.$$

Note that under our sign assumptions, the denominator is necessarily positive. The numerator defines a function $g(\lambda)$, a convex function (a parabola) with the properties

$$g'(1) \;=\; (\beta-1) - \beta\sigma\phi_y - \kappa\sigma \;<\; 0$$

and

$$g(1) \;=\; \kappa\sigma\left(\phi_\pi + \frac{(1-\beta)}{\kappa}\phi_y - 1\right),$$

so that $g(1) > 0$ if and only if the Taylor Principle is satisfied. Hence the function $g(\lambda) > 0$ for all $\lambda \leq 1$, with the consequence that $\mathrm{Det}[A(\lambda)M - I] \;>\; 0$ for all $|\lambda| \leq 1$, if and only if the Taylor Principle is satisfied.

The trace of the matrix is given by

$$\mathrm{Tr}(A(\lambda)M - I) \;=\; -\frac{1}{\Delta}\left(\frac{\Delta-\lambda(1+\sigma\kappa)}{1-\beta\lambda} + \frac{\Delta-\beta\lambda(1+\sigma\phi_y)}{1-\alpha\beta\lambda}\right).$$

The denominators of both terms inside the parentheses are positive for all $|\lambda| \leq 1$, and we necessarily have $\Delta > 0$ under our sign assumptions as well. The numerator of the first term inside the parentheses is also positive, since

$$\Delta - \lambda(1+\sigma\kappa) \;=\; \sigma\left[\kappa\phi_\pi + \phi_y - \kappa\right] \;+\; (1-\lambda)(1+\sigma\kappa) \;\geq \sigma\left[\kappa\phi_\pi + \phi_y - \kappa\right] \;>\; 0$$

if the Taylor Principle is satisfied, again using (A.6). And the numerator of the second term inside the parentheses is positive as well, since

$$\Delta - \beta\lambda(1+\sigma\phi_y) \;=\; (1-\beta\lambda)(1+\sigma\phi_y) \;+\; \kappa\sigma\phi_\pi \;>\; 0$$

under our sign assumptions. Thus the Taylor Principle is also a sufficient condition for $\mathrm{Tr}[A(\lambda)M - I] \;<\; 0$ for all $|\lambda| \leq 1$.

It then follows that the Taylor Principle is necessary and sufficient for both eigenvalues of the matrix $A(\lambda)M - I$ to have negative real part, in the case of any $|\lambda| < 1$. We use this result in the proof of Proposition 1.

A.5 Properties of the Matrix B

Necessary and sufficient conditions for both eigenvalues of a 2×2 matrix B to have modulus less than 1 are that (i) $\text{Det}B < 1$; (ii) $\text{Det}B + \text{Tr}B > -1$; and (iii) $\text{Det}B - \text{Tr}B > -1$. In the case of the matrix B defined above, we observe that

$$\Delta \, \text{Det}B = \beta, \tag{A.7}$$

$$\Delta \, \text{Tr}B = 1 + \kappa\sigma + \beta(1 + \sigma\phi_y).$$

From these facts we observe that our general sign assumptions imply that

$$\Delta \, \text{Det}B < \Delta,$$

$$\Delta \, (\text{Det}B + \text{Tr}B + 1) > 0.$$

Thus (since Δ is positive) conditions (i) and (ii) from the previous paragraph necessarily hold. We also find that

$$\Delta \, (\text{Det}B - \text{Tr}B + 1) = \kappa\sigma \left[\phi_\pi + \left(\frac{1-\beta}{\kappa} \right) \phi_y - 1 \right],$$

from which it follows that condition (iii) is also satisfied if and only if the quantity in the square brackets is positive. Thus we conclude that both eigenvalues of B have modulus less than 1 if and only if the Taylor Principle (2.14) is satisfied.

In the case that the Taylor Principle is violated (as in the case of a fixed interest rate, in which case $\phi_\pi = \phi_y = 0$), since $\text{Det}B = \mu_1\mu_2$ and $\text{Tr}B = \mu_1 + \mu_2$, where (μ_1, μ_2) are the two eigenvalues of B, the fact that condition (iii) fails to hold implies that

$$(\mu_1 - 1)(\mu_2 - 1) < 0. \tag{A.8}$$

This condition is inconsistent with the eigenvalues being a pair of complex conjugates, so in this case there must be two real eigenvalues. Condition (A.8) further implies that one must be greater than 1, while the other is less than 1. Condition (A.7) implies that $\text{Det}B > 0$, which requires that the two real eigenvalues both be non-zero and of the same sign; hence both must be positive. Thus when the Taylor Principle is violated (i.e., the quantity in (A.5) is negative), there are two real eigenvalues satisfying

$$0 < \mu_1 < 1 < \mu_2,$$

as asserted in section 2.2.

We further note that in this case, e_2', the (real) left eigenvector associated with eigenvalue μ_2, must be such that $e_2'b \neq 0$ (a result that is relied upon in section 4.2). The vector $v_2' \neq 0$ must satisfy

$$e_2' \, [B - \mu_2 I] = 0$$

to be a left eigenvector. The first column of this relation implies that $(1-\mu_2)e_{2,1} + \kappa e_{2,2} = 0$, where we use the notation $e_{2,j}$ for the jth element of eigenvector e_2'. Since $\kappa > 0$ and $\mu_2 > 1$, this requires that $e_{2,1}$ and $e_{2,2}$ must both be non-zero and have the same sign. But since both elements of b have the same sign, this implies that $e_2' b \neq 0$.

Finally, we note that whenever (A.5) holds, regardless of the sign, the eigenvalues must satisfy

$$(\mu_1 - 1)(\mu_2 - 1) \neq 0,$$

so that B has no eigenvalue equal exactly to 1. This means that the matrix $B - I$ must be non-singular, which is the condition needed for existence of unique steady-state levels of output and inflation consistent with a PFE. In the case of constant fundamentals $\omega_t = \bar{\omega}$ for all t, the unique steady-state solution to (A.2) is then given by $x_t = \bar{x}$ for all t, where

$$\bar{x} \equiv (I - B)^{-1} b\,[(1-\beta)\sigma^{-1}\bar{g} - \bar{\imath}]. \tag{A.9}$$

Note that condition (A.5) is also the condition under which $M - I$ is non-singular, as shown above. Moreover, since $I - \Lambda$ is non-singular, $M - I$ is non-singular if and only if $(I-\Lambda)(M-I) = [(I-\Lambda)M + \Lambda] - I$ is non-singular. This is the condition under which equation (A.3) has a unique steady-state solution, in which $e_t = \bar{e}$ for all t, with

$$\bar{e} \equiv (I - M)^{-1} m\,\bar{\omega}.$$

This solution for steady-state PFE expectations is consistent with (A.9) because of the identities linking the M and B matrices noted above.

A.6 Convergence of the PFE Dynamics

As noted in the text in section 2.2, in the case that $\phi_\pi = \phi_y = 0$, there exists a continuum of PFE solutions that remain bounded for all t, described by equations (2.16) for alternative values of the coefficient χ. Here we show that if after some finite date T, both $\bar{\imath}_t$ and ρ_t take constant values, then each of this continuum of solutions has the property that

$$\lim_{t\to\infty} \pi_t = \pi_{LR}, \qquad \lim_{y\to\infty} y_t = y_{LR},$$

where the limiting values are independent of χ and are given by (2.17). Moreover, the limiting values to which the PFE dynamics converge correspond to the PFE steady state (A.9).

If $\bar{\imath}_t = \bar{\imath}_{LR}$ and $\rho_t = \rho_{LR}$ for all $t \geq T$, then for any $t \geq T$, (2.16) takes the form

$$\begin{aligned} x_t &= v_1(e_1'b)\,\frac{\rho_{LR} - \bar{\imath}_{LR}}{1-\mu_1} \;-\; v_2(e_2'b)\,\left\{\sum_{j=1}^{t-T} \mu_2^{-j}\cdot(\rho_{LR}-\bar{\imath}_{LR}) \;+\; \sum_{j=t+1-T}^{t} \mu_2^{-j}(\rho_{t-j}-\bar{\imath}_{t-j})\right\} \\ &\qquad + \chi\, v_2 \mu_2^{-t} \\ &= \left[\frac{v_1(e_1'b)}{1-\mu_1} \;-\; \frac{v_2(e_2'b)}{\mu_2 - 1}(1-\mu_2^{T-t})\right]\cdot(\rho_{LR}-\bar{\imath}_{LR}) \;+\; C\, v_2\mu_2^{-t} \;+\; \chi\, v_2\mu_2^{-t} \end{aligned}$$

where

$$C \;\equiv\; \sum_{s=0}^{T-1} \mu_2^s(\rho_s - \bar{\imath}_s)$$

has a value independent of t. Given that $0 < \mu_2^{-1} < 1$, we see immediately from this that x_t converges to

$$x_{LR} \equiv \left[\frac{v_1(e_1'b)}{1-\mu_1} \;+\; \frac{v_2(e_2'b)}{\mu_2 - 1}\right]\cdot(\rho_{LR}-\bar{\imath}_{LR})$$

as $t \to \infty$. This limiting vector is independent of the value of χ.

Finally, we note that

$$\begin{aligned} (I-B)\,x_{LR} &= (I-B)\,\left[\frac{v_1(e_1'b)}{1-\mu_1} \;-\; \frac{v_2(e_2'b)}{\mu_2-1}\right]\cdot(\rho_{LR}-\bar{\imath}_{LR}) \\ &= \left[\frac{(1-\mu_1)v_1(e_1'b)}{1-\mu_1} \;-\; \frac{(1-\mu_2)v_2(e_2'b)}{\mu_2-1}\right]\cdot(\rho_{LR}-\bar{\imath}_{LR}) \\ &= [v_1(e_1'b) \;+\; v_2(e_2'b)]\cdot(\rho_{LR}-\bar{\imath}_{LR}) \\ &= b\cdot(\rho_{LR}-\bar{\imath}_{LR}), \end{aligned}$$

so that x_{LR} is just the vector of steady-state values defined in (A.9). Our definitions of B and b above further imply that when $\phi_\pi = \phi_y = 0$,

$$(I-B)^{-1}\,b \;=\; \begin{bmatrix} -\frac{1-\beta}{\kappa} \\ -1 \end{bmatrix},$$

so that (A.9) implies the values given in (2.17).

B Proofs of Propositions

B.1 Proof of Proposition 1

As discussed in the text, under the hypotheses of the proposition, there must exist a date $\bar{T}$ such that the fundamental disturbances $\{\omega_t\}$ can be written in the form

$$\omega_t \;=\; \omega_\infty + \sum_{k=1}^{K} a_{\omega,k}\lambda_k^{t-\bar{T}}$$

for all $t \geq \bar{T}$, and the initial conjecture can also be written in the form

$$e_t(0) \;=\; e_\infty(0) + \sum_{k=1}^{K} a_{e,k}(0)\lambda_k^{t-\bar{T}}$$

for all $t \geq \bar{T}$, where $|\lambda_k| < 1$ for all $k = 1, \ldots, K$. (There is no loss of generality in using the same date $\bar{T}$ and the same finite set of convergence rates $\{\lambda_k\}$ in both expressions.) With a driving process and initial condition of this special form, the solution to the system of differential equations (2.21) will be of the form

$$e_t(n) \;=\; e_\infty(n) + \sum_{k=1}^{K} a_{e,k}(n)\lambda_k^{t-\bar{T}}$$

for all $t \geq \bar{T}$, for each $n \geq 0$. We then need simply determine the evolution as n increases of the finite set of values $e_t(n)$ for $0 \leq t \leq T-1$, together with the finite set of coefficients $e_\infty(n)$ and $a_{e,k}(n)$. This is a set of $2(\bar{T}+K+1)$ functions of n, which we write as the vector-valued function $\mathbf{e}(n)$ in the text.

In the case of any belief sequences and disturbances of the form assumed in the above paragraph, it follows from (2.20) that the implied correct beliefs will be of the form

$$e_t^*(n) \;=\; e_\infty^*(n) + \sum_{k=1}^{K} a_{e,k}^*(n)\lambda_k^{t-\bar{T}}$$

for all $t \geq \bar{T}$, where

$$e_\infty^*(n) \;=\; M\, e_\infty(n) \;+\; m\,\omega_\infty,$$

and

$$a_{e,k}^*(n) \;=\; A(\lambda_k)\,[M\, a_{e,k}(n) \;+\; m\, a_{\omega,k}]$$

for each $k = 1, \ldots, K$. We further observe that for any $t < \bar{T}$,

$$\begin{aligned} e_t^*(n) \;=\; \sum_{j=1}^{\bar{T}-t-1} [\psi_j e_{t+j}(n) \;+\; \varphi_j \omega_{t+j}] \;\;+\;\; & \Lambda^{\bar{T}-t-1}\,[M e_\infty(n) \;+\; m\omega_\infty] \\ +\;\; & \sum_{k=1}^{K} \lambda_k^{-1}\Lambda^{\bar{T}-t-1}\,A(\lambda_k)\,[M a_{e,k}(n) \;+\; m a_{\omega,k}]. \end{aligned}$$

Thus the sequence $\{e_t^*(n)\}$ can also be summarized by a set of $2(\bar{T}+K+1)$ functions of n, and each of these is a linear function of the elements of the vectors $\mathbf{e}(n)$ and $\boldsymbol{\omega}$.

It then follows that the dynamics (2.21) can be written in the more compact form

$$\dot{\mathbf{e}}(n) \;=\; V\,\mathbf{e}(n) \;+ W\,\boldsymbol{\omega}, \tag{B.10}$$

where the elements of the matrices V and W are given by the coefficients of the equations in the previous paragraph. Suppose that we order the elements of $\mathbf{e}(n)$ as follows: the first two elements are the elements of e_0, the next two elements are the elements of e_1, and so on, through the elements of $e_{\bar{T}-1}$; the next two elements are the elements of $a_{e,1}$, the two elements after that are the elements of $a_{e,2}$, and so on, through the elements of $a_{e,K}$; and the final two elements are the elements of e_∞. Then we observe that the matrix V is of the form

$$V = \begin{bmatrix} V_{11} & V_{12} \\ 0 & V_{22} \end{bmatrix}, \tag{B.11}$$

where the first $2\bar{T}$ rows are partitioned from the last $2(K+1)$ rows, and the columns are similarly partitioned.

Moreover, the block V_{11} of the matrix is of the block upper-triangular form

$$V_{11} = \begin{bmatrix} -I & v_{12} & \cdots & v_{1,\bar{T}-1} & v_{1,\bar{T}} \\ 0 & -I & \cdots & v_{2,\bar{T}-1} & v_{2,\bar{T}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -I & v_{\bar{T}-1,\bar{T}} \\ 0 & 0 & \cdots & 0 & -I \end{bmatrix}, \tag{B.12}$$

where now each block of the matrix is 2×2. Furthermore, when V_{22} is similarly partitioned into 2×2 blocks, it takes the block-diagonal form

$$V_{22} = \begin{bmatrix} A(\lambda_1)M - I & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & A(\lambda_K)M - I & 0 \\ 0 & \cdots & 0 & M - I \end{bmatrix}. \tag{B.13}$$

These results allow us to determine the eigenvalues of V. The block-triangular form (B.11) implies that the eigenvalues of V consist of the $2\bar{T}$ eigenvalues of V_{11} and the $2(K+1)$ eigenalues of V_{22} (the two diagonal blocks). Similarly, the block-triangular form (B.12) implies that the eigenvalues of V_{11} consist of the eigenvalues of the diagonal blocks (each of which is $-I$), which means that the eigenvalue -1 is repeated $2\bar{T}$ times. Finally, the block-diagonal form (B.13) implies that the eigenvalues of V_{22} consist of the eigenvalues of the diagonal blocks: the two eigenvalues of $A(\lambda_k)M - I$, for each $k = 1, \ldots, K$, and the two eigenvalues of $M - I$.

Using the results in section A.3, it follows from the hypothesis that the reaction function coefficients satisfy (2.14) and the hypothesis that $|\lambda_k| < 1$ for each k that all of the eigenvalues of $M - I$ and of each of the matrices $A(\lambda_k)M - I$ have negative real part. Since all of the other eigenvalues of V are equal to -1, all $2(\bar{T} + K + 1)$ eigenvalues of V have negative real part. This implies that V is non-singular, so that

there is a unique rest point for the dynamics (B.10), defined by (3.3) in the text. It also implies that the dynamics (B.10) converge asymptotically to that rest point as n goes to infinity, for any initial condition $\mathbf{e}(0)$ (Hirsch and Smale, 1974, pp. 90-95).[77]

The rest point to which $\mathbf{e}(n)$ converges is easily seen to correspond to the unique PFE that belongs to the same linear space L^2. Beliefs in L^2 constitute a PFE if and only if $\mathbf{e}^* = \mathbf{e}$. From our characterization above of $\mathbf{e}^*$, this is equivalent to the requirement that $V\,\mathbf{e} + W = 0$, which holds if and only if $\mathbf{e} = \mathbf{e}^{PF}$, the unique rest point of the system (B.10).

Finally, the paths of output and inflation in any reflective equilibrium are given by (A.1), given the solution for $\{e_t(n)\}$. Using (2.8), one obtains a similar linear equation for the nominal interest rate each period. It then follows that for any t, the reflective equilibrium values for y_t, π_t, and i_t converge to the FS-PFE values as n is made large. Furthermore, the complete sequences of values for these three variables for any value of n depend on only the finite number of elements of the vector $\mathbf{e}(n)$, in such a way that for any $\epsilon > 0$, there exists an $\tilde{\epsilon} > 0$ such that it is guaranteed that each of the variables y_t, π_t, and i_t are within distance ϵ of their FS-PFE values for all t as long as $|\mathbf{e}(n) - \mathbf{e}^{PF}| < \tilde{\epsilon}$. The convergence of $\mathbf{e}(n)$ to $\mathbf{e}^{PF}$ then implies the existence of a finite $n(\epsilon)$ for which the latter condition is satisfied, regardless of how small $\tilde{\epsilon}$ needs to be. This proves the proposition.

B.2 Comparison with a Discrete Model of Belief Revision

Here we note that the convergence result in Proposition 1 would not hold with the same generality were we instead to assume a discrete model of belief revision in which, instead of the continuous model of belief revision (2.21), we iterate the mapping

$$e_t(N+1) \;=\; e_t^*(N) \tag{B.14}$$

for $N = 0, 1, 2, \ldots$, where for each N, $\{e_t^*(N)\}$ is the sequence of correct beliefs implied by average expectations specified by the sequence $\{e_t(N)\}$. As with the continuous model, we might take as given some "naive" initial conjecture, and then consider how it evolves as a result of further iterations of the mapping. And as with the continuous model, *if* the process converges to a fixed point, such a fixed point must correspond to PFE beliefs.

[77] Of course, it is important to recognize that this result only establishes convergence for initial conjectures that belong to the linear space L^2. The result also only establishes convergence under the assumption that the linear dynamics (B.10) apply at all times; this depends on assuming that the reaction function (2.8) can be implemented at all times, which requires that the zero lower bound never bind. Thus we only establish convergence for all those initial conjectures such that the dynamics implied by (2.21) never cause the zero lower bound to bind. There is however a large set of initial conditions for which this is true, given that the unconstrained dynamics are asymptotically convergent.

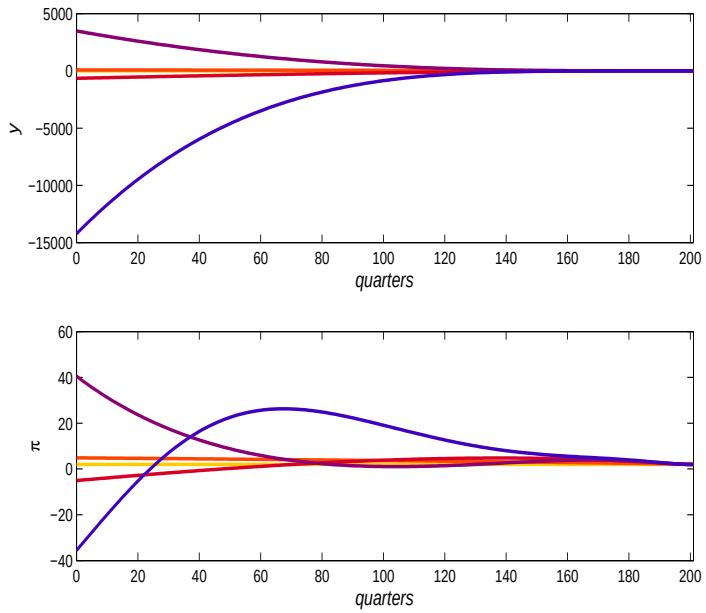

Figure 7: Reflective equilibrium outcomes for $N = 0$ through 4 (progressively darker lines) when the Taylor-rule intercept is reduced for 200 quarters, as in Figure 2; but a discrete process of iterative belief revision is assumed.

However, the conditions for convergence of the discrete process, while related to the conditions under which the continuous process converges, are more stringent. Convergence need not obtain under the conditions hypothesized in Proposition 1, as the following numerical example illustrates. In Figure 7, the same policy experiment is considered as in Figure 2, namely, the intercept of the Taylor rule is expected to be lowered for 200 quarters, after which it is expected to return to the level consistent with the inflation target π^*. All model parameters are also the same as in Figure 2, and the initial conjecture is assumed to be $e_t(0) = 0$ for all t, also as in the Figure 2. However, in Figure 7 the iterative model of belief revision (B.14) is assumed, whereas the continuous model (2.21) is assumed in Figure 2.

The figure plots the implied TE dynamics of output and inflation for iterations $N = 0, 1, 2, 3$, and 4. The belief-revision dynamics are seen to be explosive. The first revision of the initial conjecture (which takes account of the fact that it is predictable that if people maintain the initial beliefs, consistent with the unperturbed steady state, the temporary policy will lead to higher inflation and output) raises both output and inflation further. But anticipation of *these* effects (and the associated increase in the interest rate that they must provoke) should actually lead output and inflation to be *lower* in stage $N = 2$. Anticipation of the $N = 2$ outcomes (which imply an even deeper cut in the interest rate) then leads output and inflation to

be *high* again in stage $N = 3$, and to an even greater extent than in stage $N = 1$. Anticipating of this then leads output and inflation to be *low* again in stage $N = 4$, to an even greater extent than in stage $N = 2$. The oscillations continue, growing larger and larger, as N is increased; but as the figure shows, the predicted expectations are already very extreme after only four iterations of the belief updating mapping.

It is not accidental that the unstable dynamics of belief revision in this case are oscillatory. In terms of the compact notation introduced in the proof of Proposition 1 (under the assumption of exponentially convergent fundamentals and average beliefs), the discrete model of belief revision (B.14) replaces the continuous dynamics (B.16) by the discrete process

$$\mathbf{e}(N+1) \;=\; (I+V)\,\mathbf{e}(N) \;+W\,\boldsymbol{\omega}. \tag{B.15}$$

This process is unstable if not all eigenvalues of $I + V$ are of modulus less than 1. Since the eigenvalues of $I + V$ are equal to $1 + \mu_i$, where μ_i is an eigenvalue of V, and we have shown above that all eigenvalues of V have negative real part, $I + V$ cannot have a real eigenvalue greater than 1. It can, however, have a real eigenvalue with *modulus* greater than 1, if V has a real eigenvalue that is less than -2. This is the case shown in Figure 7, in which a large negative eigenvalue results in explosive oscillations.

We feel, however, that the kind of unstable process of belief revision illustrated by Figure 7 is unrealistic, as it is requires that at each stage in the reasoning, one must conjecture that *everyone* else should reason in one precise way, even though that assumed reasoning changes dramatically from each stage in the process of reflection to the next. The continuous process of belief revision proposed in the text avoids making such an implausible assumption.

B.3 Proof of Proposition 2

It has already been shown in the text that under the assumptions of the proposition, we have $e_t(n) = e_{LR}(n)$ for all $t \geq T$, where $e_{LR}(n)$ is given by (3.6). It has also been shown that for any $\tau \geq 1$, the solution for $e_\tau(n)$, where $\tau \equiv T - t$ is the number of periods remaining until the regime change, is independent of T. The functions $\{e_\tau(n)\}$ further satisfy the system of differential equations

$$\begin{aligned}\dot{e}_\tau(n) \;=\; -e_\tau(n) \;\;+\;\; (I-\Lambda)\sum_{j=1}^{\tau-1}\Lambda^{j-1}\left[Me_{\tau-j}(n) \,+\, m_2\bar{\imath}_{SR}\right] \\ +\;\; \Lambda^{\tau-1}\left[Me_{LR}(n) \,+\, m_2\bar{\imath}_{LR}\right]\end{aligned} \tag{B.16}$$

derived in the text, together with the initial conditions $e_\tau(0) = 0$ for all $\tau \geq 1$. (Equation (B.16) repeats equation (3.9) from the text.)

We wish to calculate the behavior of the solution to this system as $\tau \to \infty$ for an arbitrary value of n. It is convenient to use the method of z-transforms (Jury, 1964). For any n, let the z-transform of the sequence $\{e_\tau(n)\}$ for $\tau \geq 1$ be defined as the function

$$X_n(z) \;\equiv\; \sum_{\tau=1}^{\infty} e_\tau(n)\, z^{1-\tau}. \tag{B.17}$$

Here $X_n(z)$ is a vector-valued function; each element is a function of the complex number z, defined for complex numbers $|z| > 1/\rho$, where ρ is the minimum of the radii of the convergence of the two series.

Differentiating (B.17) with respect to n, and substituting (B.16) for $\dot{e}_\tau(n)$ in the resulting equation, we obtain an evolution equation for the z-transform:

$$\begin{aligned}
\dot{X}_n(z) \;&=\; -\sum_{\tau=1}^{\infty} e_\tau(n) z^{1-\tau} \;+\; (I-\Lambda)\sum_{j=0}^{\infty}\Lambda^j z^{-j}\left[M\sum_{\tau=1}^{\infty} e_\tau(n) z^{-\tau} \;+\; m_2\bar{\imath}_{SR}\sum_{\tau=1}^{\infty} z^{-\tau}\right] \\
&\qquad\qquad +\sum_{j=0}^{\infty}\Lambda^j z^{-j}\left[M\, e_{LR}(n) \;+\; m_2\bar{\imath}_{LR}\right] \\
&=\; -X_n(z) \;+\; (I-\Lambda)(I-\Lambda z^{-1})^{-1}\left[z^{-1}M\, X_n(z) \;+\; (z-1)^{-1}\, m_2\bar{\imath}_{SR}\right] \\
&\qquad\qquad +(I-\Lambda z^{-1})^{-1}\left[M e_{LR}(n) \;+\; m_2\bar{\imath}_{LR}\right],
\end{aligned} \tag{B.18}$$

which holds for any $n > 0$ and any z in the region of convergence. We note that the right-hand side of (B.18) is well-defined for all $|z| > 1$.

The z-transform of the initial condition is simply $X_0(z) = 0$ for all z. Thus we wish to find functions $\{X_n(z)\}$ for all $n \geq 0$, each defined on the region $|z| > 1$, that satisfy (B.18) for all n and all $|z| > 1$, together with the initial condition $X_0(z) = 0$ for all z. If we can find such a solution, then for any n we can find the implied sequence $\{e_t(n)\}$ by inverse z-transformation of the function $X_n(z)$.

We note that the dynamics of $X_n(z)$ implied by (B.18) is independent for each value of z. (This is the advantage of z-transformation of the original system of equations (B.16).) Thus for each value of z such that $|z| > 1$, we have an independent first-order ordinary differential equation to solve for $X_n(z)$, with the single initial condition $X_0(z) = 0$. This equation has a closed-form solution for each z, given by

$$\begin{aligned}
X_n(z) \;&=\; (1-z^{-1})^{-1}\left[I \;-\; \exp(n(M-I))\right](I-M)^{-1}\cdot m_2\bar{\imath}_{LR} \\
&\qquad +(z-1)^{-1}\left[I \;-\; \exp(-n\Phi(z))\right]\Phi(z)^{-1}(I-\Lambda)(I-\Lambda z^{-1})^{-1} \\
&\qquad\qquad\qquad \cdot m_2(\bar{\imath}_{SR}-\bar{\imath}_{LR})
\end{aligned} \tag{B.19}$$

for all $n \geq 0$, where

$$\Phi(z) \;\equiv\; I \;-\; (I-\Lambda)(I-\Lambda z^{-1})^{-1} z^{-1}\, M.$$

Note also that the expression on the right-hand side of (B.19) is an analytic function of z everywhere in the complex plane outside the unit circle, and can be expressed as a sum of powers of z^{-1} that converges everywhere in that region. Such a series expansion of $X_n(z)$ for any n allows us to recover the series of coefficients $\{e_\tau(n)\}$ associated with the reflective equilibrium with degree of reflection n.

For any value of $n \geq 0$, we are interested in computing

$$e_{SR}(n) \equiv \lim_{T\to\infty} e_t(n) = \lim_{\tau\to\infty} e_\tau(n).$$

The final value theorem for z-transforms[78] implies that

$$\lim_{\tau\to\infty} e_\tau(n) = \lim_{z\to 1}(z-1)X_n(z)$$

if the limit on the right-hand side exists. In the case of the solution (B.19), we observe that the limit is well-defined, and equal to

$$\lim_{z\to 1}(z-1)X_n(z) = [I - \exp(n(M-I))]\,(I-M)^{-1}\,m_2\bar{\imath}_{SR}.$$

Hence for any t and any n, $e_t(n)$ converges to a well-defined (finite) limit as T is made large, and the limit is the one given in the statement of the proposition.

B.4 Proof of Proposition 3

The result that

$$\lim_{T\to\infty} e_t(n) = e_{SR}(n)$$

for all t and n follows from Proposition 2. If in addition, the Taylor Principle (2.14) is satisfied, then as shown in section A.3 above, both eigenvalues of $M-I$ have negative real part. From this (3.8) follows; substituting of this into (3.10) yields

$$\lim_{n\to\infty} e_{SR}(n) = \bar{e}_{SR}^{PF},$$

where $\bar{e}_{SR}^{PF}$ is defined in (3.11). This establishes the first double limit in the statement of the proposition.

The result that

$$\lim_{n\to\infty} e_t(n) = e_t^{PF}$$

for all t follows from Proposition 1. Establishing the second double limit thus requires us to consider how e_t^{PF} changes as T is made large.

As discussed in section A.2 above, the FS-PFE dynamics $\{e_t^{PF}\}$ satisfy equation (A.3) for all t. Under the kind of regime assumed in this proposition (with ω_t equal

[78] See, for example, Jury (1964), p. 6.

to a constant vector $\bar{\omega}$ for all $t \geq T$), the FS-PFE (obtained by "solving forward" the difference equation) involves a constant vector of expectations, $e_t^{PF} = \bar{e}_{LR}^{PF}$ for all $t \geq T-1$, where

$$\bar{e}_{LR}^{PF} \equiv [I - M]^{-1} m_2 \bar{\imath}_{LR}$$

is the same as the vector defined in (3.7).

For periods $t < T-1$, one must instead solve the difference equation backward from the terminal condition $e_{T-1}^{PF} = \bar{e}_{LR}^{PF}$. We thus obtain a difference equation of the form

$$e_\tau \;=\; [(I-\Lambda)M \;+\; \Lambda]\, e_{\tau-1} \;+\; (I-\Lambda)\, m_2 \bar{\imath}_{SR} \tag{B.20}$$

for all $\tau \geq 2$, with initial condition $e_1 = \bar{e}_{LR}^{PF}$. The asymptotic behavior of these dynamics as τ is made large depends on the eigenvalues of the matrix

$$(I-\Lambda)M \;+\; \Lambda \;=\; C^{-1}BC, \tag{B.21}$$

which must be the same as the eigenvalues of B. (Note that (B.21) follows from (A.4).)

Under the hypothesis that the response coefficients satisfy the Taylor Principle (2.14), both eigenvalues of B are inside the unit circle. It then follows that the dynamics (B.20) converge as $\tau \to \infty$ to the steady-state vector of expectations $\bar{e}_{SR}^{PF}$ defined in (3.11). We thus conclude that

$$\lim_{T \to \infty} e_t^{PF} \;=\; \bar{e}_{SR}^{PF}$$

for any t. This establishes the second double limit.

B.5 Proof of Proposition 4

The proof of this proposition follows exactly the same lines as the proof of Proposition 1. While the definition of the matrices of coefficients V and W must be modified, it continues to be possible to write the belief revision dynamics in the compact form (B.10), for an appropriate definition of these matrices. (This depends on the fact that we have chosen $\bar{T} \geq T$, so that the coefficients of the monetary policy reaction function do not change over time during periods $t \geq \bar{T}$. Variation over time in the reaction function coefficients does not prevent us from writing the dynamics in the compact form, as long as it occurs only prior to date $\bar{T}$; and our method of analysis requires only that $\bar{T}$ be finite.)

Moreover, it continues to be the case that V will have the block-triangular form indicated in equations (B.11)–(B.13). In equation (B.13), the matrix M is defined using the coefficients (ϕ_π, ϕ_y) that apply after date T, and thus that satisfy the Taylor Principle (2.14), according to the hypotheses of the proposition. The eigenvalues of V again consist of -1 (repeated $2\bar{T}$ times); the eigenvalues of $A(\lambda_k)M$, for $k = 1, \ldots, K$,

and the eigenvalues of M. Because M is defined using coefficients that satisfy the Taylor Principle, we again find that all of the eigenvalues of M and of $A(\lambda_k)M$ have negative real part. Hence all of the eigenvalues of V have negative real part. This again implies that the dynamics (B.10) are asymptotically stable, and the fixed point to which they converge again corresponds to the FS-PFE expectations. This establishes the proposition.

Note that this result depends on the hypothesis that from date T onward, monetary policy is determined by a reaction function with coefficients that satisfy the Taylor Principle. If we assumed instead (as in the case emphasized in Cochrane, 2015a) that after date T, policy again consists of a fixed interest rate, but one that is consistent with the long-run inflation target (i.e., $\bar{\imath}_{LR} = 0$), the belief-revision dynamics would *not* converge. (See the discussion in section 4.3 of the text of the case in which an interest-rate peg differs temporarily from the long-run interest-rate peg.)

If the interest rate is also fixed after date T (albeit at some level $\bar{\imath}_{LR} \neq \bar{\imath}_{SR}$), the belief-revision dynamics can again be written in the compact form (B.10), and the matrix V will again have the form (B.11)–(B.13). But in this case, the matrix M in (B.13) would be defined using the response coefficients $\phi_\pi = \phi_y = 0$, so that the Taylor Principle is violated. It then follows from our results above that M will have a positive real eigenvalue. (By continuity, one can show that $A(\lambda_k)M$ will also have a positive real eigenvalue for all values of λ_k near enough to 1.) Hence V will have at least one (and possibly several) eigenvalues with positive real part, and the belief-revision dynamics (B.10) will be explosive in the case of almost all initial conjectures (even restricting our attention to conjectures within the specified finite-dimensional family).

B.6 Proof of Proposition 5

The proof of this proposition follows similar lines as the proof of Proposition 2. In general, the characterization of reflective equilibrium is more complex when the monetary policy response coefficients are not time-invariant, as in the situation considered here. However, in the case hypothesized in the proposition, $g_t = 0$ and from period T onward, monetary policy is consistent with constant inflation at the rate π^*. Under these circumstances, and initial conjecture under which $e_t = 0$ for all $t \geq T$ implies correct beliefs $e_t^* = 0$ for all $t \geq T$ as well. Hence under the belief-revision dynamics, the conjectured beliefs are never revised, and $e_t(n) = 0$ for all degrees of reflection $n \geq 0$, and any $t \geq T$. This result would be *the same* if we were to assume a fixed interest rate for all $t \geq T$ (that is, if we were to assume response coefficients $\phi_\pi = \phi_y = 0$ after date T, just like we do for dates prior to T), but a fixed interest rate $\bar{\imath}_t = 0$ for all $t \geq T$ (that is, the fixed interest rate consistent with the steady state with inflation rate π^*).

Thus the reflective equilibrium is the same (in this very special case) as if we

assumed a fixed interest rate in all periods (and thus the same response coefficients in all periods), but $\bar{\imath}_t = \bar{\imath}_{SR}$ for $t < T$ while $\bar{\imath}_t = 0$ for $t \geq T$.[79] And the latter is a case to which Proposition 2 applies. (Note that Proposition 2 requires no assumptions about the response coefficients except that they are constant over time, and that they satisfy (A.5). Hence the case in which $\phi_\pi = \phi_y = 0$ in all periods is consistent with the hypotheses of that proposition.)

Proposition 2 can then be used to show that the reflective equilibrium beliefs $\{e_t(n)\}$ for any degree of reflection n converge to a well-defined limiting value $e_{SR}(n)$, which is given by (3.10)–(3.11). This establishes the proposition.

B.7 Proof of Proposition 6

Let $\{e_t^1\}$ be the sequence of expectations in a reflective equilibrium when the date of the regime change is T, and $\{e_t^2\}$ be the expectations in the equilibrium corresponding to the same degree of reflection n when the date of the regime change is $T' > T$. Similarly, let $\{a_t^1\}$ and $\{a_t^2\}$ be the evolution of the vectors of summary variables that decisionmakers need to forecast in the two equilibria, and $\{e_t^{*1}\}$ and $\{e_t^{*2}\}$ the implied sequences of correct forecasts in the two equilibria. We similarly use the notation $M^{(i)}, m^{(i)}, C^{(i)}, c^{(i)}$ to refer to the matrices M, m, C, and c respectively, defined using the monetary policy response coefficients associated with regime i (for $i = 1, 2$).[80]

Let us first consider the predictions regarding reflective equilibrium in periods $t \geq T'$. Under both of the assumptions about policy, policy is expected to be the same at all dates $t \geq T'$. Since it is assumed that we start from the same initial conjecture $\{e_t(0)\}$ in both cases, and the model is purely forward-looking, it follows that the belief-revision dynamics will also be the same for all $t \geq T'$ in both cases. Hence we obtain the same sequences $\{e_t(n)\}$ in both cases, for all $t \geq T'$; and since the outcomes for output and inflation are then given by (A.1), these are the same for all $t \geq T'$ as well. Moreover, it is easily shown that under our assumptions, the common solution is one in which $e_t(n) = 0$ for all $t \geq T'$, and correspondingly $x_t(n) = 0$ for all $t \geq T'$.

Moreover, since outcomes for output and inflation are the same for all $t \geq T'$ in the two cases, it follows that the sequences of correct forecasts $\{e_t^*\}$ are the same in both cases for all $t \geq T' - 1$. (Note that the correct forecasts in period $T' - 1$ depend only on the equilibrium outcomes in period T' and later.) Hence the belief-revision

[79] Note that these two different specifications of monetary policy would *not* lead to the same reflective equilibrium expectations, under most assumptions about the real shocks or about the initial conjecture; see the discussion at the end of the proof of Proposition 4. Here we get the same result *only* because we assume $g_t = 0$ (exactly) for all $t \geq T$ and an initial conjecture under which $e_t(0) = 0$ (exactly) for all $t \geq T$.

[80] By "regime 1" we mean the Taylor rule (the regime in place in periods $T \leq t < T'$ under policy 1); by "regime 2" we mean the interest-rate peg at $\bar{\imath}_{SR}$.

dynamics for period $T'-1$ will also be the same in both cases, and we obtain the same vector $e_{T'-1}(n)$ for all n; and again the common beliefs are $e_{T'-1}(n) = 0$.

Let us next consider reflective equilibrium in periods $T \leq t \leq T'-1$. Suppose that for some such t and some n, $e_t^2 \geq e_t^1 \geq 0$ (in both components). Then

$$a_t^2 - a_t^1 \;=\; M^{(2)}\,(e^2 - e_t^1) \;+\; [M^{(2)} - M^{(1)}]\,e_t^1 \;|+\; m_2^{(2)}\,\bar{\imath}_{SR}.$$

Moreover, we observe from the above definitions of M and m that $M^{(2)}$ is positive in all elements; $M^{(2)} - M^{(1)}$ is positive in all elements; and $m_2^{(2)}$ is negative in both elements. Under the hypotheses that $e_t^2 \geq e_t^1 \geq 0$ and $\bar{\imath}_{SR} < 0$, it follows that $a_t^2 - a_t^1 \;>>\; 0$, where we use the symbol $>>$ to indicate that the first vector is greater in both elements.

Now suppose that for some n, $e_t^2 \geq e_t^1 \geq 0$ for all $T \leq t \leq T'-1$. It follows from our conclusions above that these inequalities then must hold for all $t \geq T$. It also follows from the argument in the paragraph above that we must have $a_t^2 >> a_t^1$ for all $T \leq t \leq T'-1$, along with $a_t^2 = a_t^1$ for all $t \geq T'$. This implies that $e_t^{*2}(n) >> e_t^{*1}(n)$ for all $T \leq t < T'-1$, while $e_t^{*2}(n) = e_t^{*1}(n)$ for $t = T'-1$.

The fact that $e_t^{*2}(n) = e_t^{*1}(n)$ for $t = T'-1$ means that the belief-revision dynamics for period $T'-1$ will again be the same in both cases, and we obtain the same vector $e_{T'-1}(n)$ for all n; and again the common beliefs are $e_{T'-1}(n) = 0$. For periods $T \leq t < T'-1$, we continue to have $e_t^{*1}(n) = 0$ for all n, for the same reason as in the case of periods $t \geq T'$. But now the fact that we start from the common initial conjecture $e_t^2(0) = e_t^1(0) = 0$ implies that $e_t^{*2}(0) >> e_t^{*1}(0) = 0$ and hence $\dot{e}_t^2(0) >> \dot{e}_t^1(0) = 0$. This implies that for small enough $n > 0$, we will have $e_t^2(n) >> e_t^1(n) = 0$ for all $T \leq t < T'-1$.

Moreover, for any n, as long as we continue to have $e_t^2(n) \geq e_t^1(n) = 0$ for all $t \geq T$, we will continue to have $e_t^{*2}(n) >> e_t^{*1}(n) = 0$ for all $T \leq t < T'-1$. Since the belief-revision dynamics (2.21) imply that for any $n > 0$, $e_t(n)$ is an average of $e_t(0)$ and the vectors $e_t^*(\tilde{n})$ for values $0 \leq \tilde{n} < n$, as long as we have had $e_t^{*2}(\tilde{n}) >> 0$ for all $0 \leq \tilde{n} < n$, we will necessarily have $e_t^2(n) >> 0$. Thus we conclude by induction that $e_t^2(n) >> e_t^1(n) = 0$ for all $n > 0$, and any $T \leq t < T'-1$.

The associated reflective equilibrium outcomes are given by (A.1) in each case. This implies that

$$x_t^2 - x_t^1 \;=\; C^{(2)}\,(e^2 - e_t^1) \;+\; [C^{(2)} - C^{(1)}]\,e_t^1 \;|+\; c_2^{(2)}\,\bar{\imath}_{SR}.$$

Note furthermore that all elements of $C^{(2)}$ are non-negative, with at least one positive element in each row; that all elements of $C^{(2)} - C^{(1)}$ are positive; and that all elements of $c_2^{(2)}$ are negative. Then the fact that $e_t^2(n) \geq e_t^1(n) = 0$ for all $T \leq t \leq T'-1$ and $\bar{\imath}_{SR} < 0$ implies that $x_t^2 \;>>\; x_t^1$ for all $T \leq t \leq T'-1$.

Finally, let us consider reflective equilibrium in periods $0 \leq t < T$. In these periods, the monetary policy is expected to be the same in both cases (the fixed

interest rate). Suppose that for some such t and some n, $e_t^2 \geq e_t^1$. Then

$$a_t^2 - a_t^1 \;=\; M^{(2)}\,(e^2 - e_t^1) \;\geq\; 0,$$

because all elements of $M^{(2)}$ are positive. Since we have already concluded above that $a_t^2 >> a_t^1$ for all $T \leq t \leq T' - 1$, and that $a_t^2 = a_t^1$ for all $t \geq T'$, this implies that $e_t^{*2} >> e_t^{*1}$ for all $0 \leq t < T$.

We can then use an inductive argument, as above, to show that $e_t^2(n) >> e_t^1(n)$ for any $n > 0$, and any $0 \leq t < T$. It follows from this that

$$x_t^2 - x_t^1 \;=\; C^{(2)}\,(e^2 - e_t^1) \;>>\; 0$$

for any $n > 0$, and any $0 \leq t < T$, given that all elements of $C^{(2)}$ are non-negative, with at least one positive element in each row. This establishes the proposition.