The Predictive Content of the Output Gap for Inflation: Resolving In–Sample and Out–of–Sample Evidence *

Todd E. Clark
Federal Reserve Bank of Kansas City

Michael W. McCracken
University of Missouri–Columbia

January 2003
[Preliminary and incomplete]

Abstract

In light of the gulf between existing in–sample and out–of–sample evidence on the Phillips curve, this paper sifts through potential explanations for the gap in the evidence, focusing on models relating inflation to the output gap. One potential explanation is simply power: the out–of–sample evidence is relatively weak because the out–of–sample metrics are less powerful than usual in–sample causality tests. A second potential explanation is model instability — shifts in the coefficients or residual variance of the inflation–output gap model. This paper evaluates these explanations on the basis of comparisons of the sample forecasting results to results from Monte Carlo simulations of DGPs that either impose stability or allow select, empirically–identified breaks in the coefficients of the DGP. This evidence shows that while power differences can account for some portion of the gulf between the in–sample and out–of–sample success of the Phillips curve, much of the gulf is attributable to instabilities in the model, particularly in the coefficients on the output gap. Theoretical analysis, based on a local alternatives framework, confirms that breaks in the output gap coefficients, but not breaks in residual variances or AR coefficients, can lead to a breakdown in the power of tests of equal forecast accuracy and forecast encompassing.

JEL Nos.: E37, E31, C53, C52
Keywords: Phillips curve, forecasts, causality, break test

* *Clark*: Economic Research Dept.; Federal Reserve Bank of Kansas City; 925 Grand; Kansas City, MO 64198; `todd.e.clark@kc.frb.org`. *McCracken*: Dept. of Economics; University of Missouri-Columbia; 118 Professional Building; Columbia, MO 65211; `mccrackenm@missouri.edu`. Jushan Bai kindly provided computer programs for computing the Bai-Perron break tests. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

1 Introduction

The Phillips curve — broadly defined here and in other studies such as Stock and Watson (1999b) as a model relating inflation to the unemployment rate, output gap, or capacity utilization — remains one of the most widely used and studied models in macroeconomics. In particular, variants of the Phillips curve continue to appear in efforts to model and understand business cycle fluctuations, forecast inflation, and analyze monetary policy.[1] Historically, while many have contended the basic model warrants a key place in macroeconomics (a recent example is Blinder (1997)), others have argued otherwise (a recent example is Rogerson (1997)). References to the considerable historical debate over the Phillips curve are provided in such summaries as Gordon (1990), King and Watson (1994), and Atkeson and Ohanian (2001). In recent years, the Phillips curve in the form of a model relating inflation to the output gap has enjoyed something of a renaissance, in part because the supply side of the standard dynamic, stochastic general equilibrium model with sticky prices reduces to an expectational form of a Phillips curve (see, for example, Rotemberg and Woodford (1997), King (2000), and McCallum and Nelson (2001)).[2]

A common metric for empirically evaluating Phillips curve models is their performance in out–of–sample forecasting. The models are, after all, often used for forecasting. In fact, some studies — examples include Kuttner (1994), St-Amant and van Norden (1997), Gerlach and Smets (1999), and Peersman and Smets (1999) — explicitly estimate or define the output gap as the component of output that has predictive power for inflation. Moreover, even if forecasting is not an explicit objective, out–of–sample forecast performance has come to be widely viewed as a useful metric for evaluating the validity of a model.[3] For example, since the work of Meese and Rogoff (1983, 1988) it has become standard in the empirical exchange rate literature to use forecast performance as a metric for evaluating models of the exchange rate; recent examples include Mark (1995), Clarida, Sarno, Taylor, and

[1] Some examples include Roberts (1995), Fisher, Mahadeva, and Whitley (1997), St-Amant and van Norden (1997), Gerlach and Smets (1999), Peersman and Smets (1999), Reifschneider, Tetlow, and Williams (1999), Rotemberg and Woodford (1997), Rudebusch and Svensson (1999), Claus (2000), and McCallum and Nelson (2001).

[2] In the more general form, the standard DSGE model implies inflation is a function of expected future inflation and current marginal cost (see, for example, Gali and Gertler (1999) and Sbordone (2002)). Under certain assumptions, the marginal cost variable can be replaced by the output gap. Although many studies implicitly or explicitly impose the necessary assumptions, some have argued the assumptions are unreasonable (see, for example, Gali, Gertler, and Lopez-Salido (2001)).

[3] Although forecast performance is often used as a general metric of model validity, some (such as Clements (2002) and Inoue and Kilian (2002)) question the practice.

Valente (2001), and Kilian and Taylor (2001). A number of Phillips curve analyses have also explicitly treated forecast performance as a test of model validity. Examples include Fisher, Mahadeva, and Whitley (1997), Atkeson and Ohanian (2001), and Camba–Mendez and Rodriguez–Palenquela (2001).

As is the case with so many relationships of interest in macroeconomics, there is a considerable body of evidence that, despite seemingly good in–sample fits, the out-of-sample forecast performance of Phillips curve models is mixed. While the generalized Phillips curve literature is far too long to be tractably surveyed here, recent examples of studies documenting a good in–sample fit of reduced form inflation–output gap specifications include Kuttner (1994), Gerlach and Smets (1999), Peersman and Smets (1999), Rudebusch and Svensson (1999), and Orphanides (2000). Of analyses that consider out–of–sample performance, some conclude that Phillips formulations generally forecast well; recent examples include Staiger, Stock, and Watson (1997), Stock and Watson (1999b), and Fisher, Liu, and Zhu (2002). Other studies reach the opposite conclusion. For example, Atkeson and Ohanian (2001) conclude that Phillips curve models using the unemployment rate and the real activity index developed by Stock and Watson (1999b) fail to out–forecast a random walk. Orphanides and van Norden (2001) show that estimates of the output gap made in real time often fail to improve forecasts from a simple autoregression.[4]

In light of the gap between the in–sample and out–of–sample evidence on the Phillips curve, in this paper we sift through potential explanations for the gap in the evidence, focusing on models relating inflation to the output gap because this particular formulation seems to have garnered the greatest interest in recent years. One potential explanation is simply power: the out–of–sample evidence is relatively weak because the out–of–sample metrics are less powerful than usual in–sample causality tests. Stressing the lower power of out–of–sample tests, Inoue and Kilian (2002) advocate the use of in–sample tests for model evaluation. Power might be an especially good explanation in the Phillips curve case to the extent that the relative power of out–of-sample metrics declines as the forecast horizon increases, because in–sample evaluation of models relating inflation to the output gap routinely use quarterly inflation rates while forecast evaluations often include longer forecast horizons.[5]

[4]Camba–Mendez and Rodriguez–Palenzuela (2001) report similar findings, though their analysis abstracts from source data revisions.

[5]The relationship of power to forecast horizon (in an absolute sense, rather than in the relative sense of out-of-sample compared to in–sample power) has been the subject of debate in the exchange rate literature.

A second potential explanation is model instability — shifts in the coefficients or residual variance of the inflation–output gap model. Stock and Watson (1996, 1999a, 2001) document pervasive instabilities in reduced–form macroeconomic relationships; Cecchetti (1995) and Stock and Watson (2001) present evidence of instabilities in forecast performance. Clark and McCracken (2002b) provide analytical and Monte Carlo evidence on how such shifts can affect tests of equal forecast accuracy and encompassing. In the specific case of the Phillips curve, though, the existing evidence suggests little role for instability.[6] Stock and Watson (1999b) report that while there is evidence of instability in the AR coefficients of an inflation–unemployment rate model, the apparent shifts don't affect forecast performance much, while the coefficients on unemployment appear stable. Rudebusch and Svensson (1999) apply stability tests to the full set of coefficients of an inflation–output gap model and are unable to reject stability.[7] Similarly, detailed test results reported in Stock and Watson (2001) show inflation–output gap models (and other Phillips curve specifications) to be largely stable.

Our results show that while power differences can account for some portion of the gulf between the in–sample and out–of–sample success of the Phillips curve, much of the gulf is attributable to instabilities in the model, particularly in the coefficients on the output gap. We arrive at this conclusion by comparing sample results for a range of inflation and output gap measures to results of Monte Carlo simulations of empirically–based data–generating processes that (i) impose stability or (ii) allow various types of discrete breaks in the model parameters. The imposed shifts in the DGPs are based on the results of extensive break test analysis applied to the empirical models in question. As we document, the break tests provide strong evidence of shifts in the residual variances of the inflation equations, modest evidence of shifts in the AR parameters of the inflation equations, and only very limited evidence of shifts in the output gap coefficients. However, simulations show that breaks in residual variances or AR parameters cannot account for the out–of–sample breakdown in the output gap's predictive power observed in the data. Simulations

See, for example, Mark (1995), Kilian (1999), and Mark and Sul (2002).

[6] A variety of studies have found that apparent instabilities in particular models can be overcome with modifications of the benchmark model. For example, Staiger, Stock, and Watson (2001) show that models relating inflation to the rate of unemployment or capacity utilization can be made stable by allowing an appropriate univariate trend in the measure of economic activity. Related examples include Lown and Rich (1997) and Brayton, Roberts, and Williams (1999).

[7] Using a simulation approach, Rudebusch (2002) finds that structural changes in policy don't seem to produce identifiable shifts in a reduced form model for the economy that includes an inflation–output gap form of the Phillips curve.

show instead that the sample forecasting results are best explained by discrete shifts in the output gap coefficients — shifts that, in most cases, the break tests have only modest power to identify. Building on the work of Clark and McCracken (2002b), we conclude the paper by showing analytically that breaks in the output gap coefficients, but not breaks in residual variances or AR coefficients, lead to a breakdown in the power of tests of equal forecast accuracy and forecast encompassing. We do so using a local alternatives framework, like that of Hansen (2000).

The paper proceeds as follows. Section 2 reviews the data and model specifications. Section 3 explains our forecast evaluation methodology, including the data sample and forecasting scheme, forecast tests, and inference approach. Section 4 presents the sample forecasting results. Section 5 uses Monte Carlo simulations to examine the ability of power differences to account for the gap between the in–sample and out–of–sample evidence. Section 6 presents our evidence on structural instabilities in the Phillips curve. Section 7 uses Monte Carlo simulations to examine the ability of the identified breaks to account for the sample results. Section 8 then presents our theoretical analysis of the effects of breaks on forecast—based tests.

2 Data and Model Specifications

2.1 Data

As previous analyses have relied on a range of inflation measures and gap definitions, we present complete results for two measures of inflation and two measures of the output gap. Inflation is measured using the GDP (chain) price index and the CPI excluding food and energy, or core CPI. The GDP price index (or, equivalently, the deflator) and CPI seem to be the most commonly used measures in the recent generalized Phillips Curve literature. We use the core CPI in lieu of the overall CPI because policymakers tend to focus on ex food and energy inflation and because using the core CPI mitigates the role of energy price shocks. Results not reported — except in the case of the structural break tests, for reasons given below — for the chain price index for consumption excluding food and energy are qualitatively similar.

For the output gap, we use the measures most common in the recent Phillips Curve literature: (1) actual GDP less the CBO's estimate of potential GDP (in logs) and (2) Hodrick and Prescott (1997)–filtered GDP. In results not reported, we also considered the output

gap computed with a one–sided version of the HP filter. Stock and Watson (1999b,2001) explicitly use the one–sided filter because it produces the gap that would be estimated in real time. While our one–sided output gap estimate yielded results that are in many respects qualitatively similar to those we report, we found the one–sided measure has less predictive power for inflation (in–sample and out–of–sample), and we were somewhat less successful in accounting for its inconsistent forecast performance. Presumably, the lower and less consistent predictive power of the one–sided gap reflects not only the greater volatility of the one–sided measure but also the added difficulty of forecasting in real time, highlighted by Orphanides and van Norden (2001).

All data except the CBO's estimate of potential GDP were obtained from the FAME database. The potential GDP series was downloaded from the Federal Reserve Bank of St. Louis' website. The data span 1957:Q1 through 2001:Q4, except that data on the core PCE price index used in some supplemental results don't begin until 1959:Q1.

2.2 Models

Following the work of Stock and Watson (1999b,2001), among many others, we treat inflation as being close enough to I(1) to warrant imposing I(1) and compare forecasts of the change in inflation from a Phillips Curve specification including the output gap to forecasts from a simple autoregressive model. While not reported, models specified in inflation *levels* yield qualitatively similar forecast results. We report forecast results for the two horizons that seem to be of most interest to policymakers and most widely used in previous studies: one quarter and four quarters.

Letting τ denote the forecast horizon (in quarters), we use reduced–form Phillips curves

$$\pi_{t+\tau}^{(\tau)} - \pi_t = \alpha_0 + \sum_{l=0}^{L-1} \alpha_l \Delta\pi_{t-l} + \sum_{m=0}^{M-1} \beta_m y_{t-m} + u_{2,t+\tau}, \tag{1}$$

where inflation is $\pi_t^{(\tau)} \equiv (400/\tau) * \ln(p_t/p_{t-\tau})$, $\pi_t^{(1)} \equiv \pi_t$, and y_t is the output gap in percentage terms (in logs multiplied by 100). The same basic form has been used in studies such as Stock and Watson (1999b,2001). In our forecast evaluation, we focus on comparing projections from this model to those from a model without the output gap. This AR specification takes the form

$$\pi_{t+\tau}^{(\tau)} - \pi_t = \alpha_0 + \sum_{l=0}^{L-1} \alpha_l \Delta\pi_{t-l} + u_{1,t+\tau}. \tag{2}$$

In results not reported in the interest of brevity, we also compare our Phillips curve forecasts against a random walk benchmark, as some inflation studies have used a random walk (no change) forecast as a benchmark (examples include Atkeson and Ohanian (2001) and Stock and Watson (1999b)).[8] We have omitted these results because they are qualitatively very similar to those reported.

For simplicity, we use the same lag lengths across forecast horizons and all time periods. In particular, for both the GDP price index and the core CPI, we always include two lags of the change in inflation ($L = 2$) and one lag of the output gap ($L = 1$).[9] Allowing two lags of inflation minimizes the AIC for full–sample estimates of the autoregression (2) for both $\tau = 1$ and $\tau = 4$. For $\tau = 1$, two lags is also always the SIC-minimizing specification. Imposing the same number of inflation lags in the Phillips curve (1), allowing just one lag of the output gap minimizes the SIC for full–sample estimates for this specification for $\tau = 1$. In most cases, one lag of the gap is also the SIC–preferred choice when $\tau = 4$. We explicitly opted to allow for a more profligate specification of inflation lags than gap lags because: (1) the optimal number of inflation lags seems to be sensitive to the sample period, and (2) most studies of inflation and output gap relationships use just one lag of the output gap. In results not reported in the interest of brevity, we allowed the lag length to be chosen at each point in time as forecasting proceeds, and obtained qualitatively similar results.

3 Forecast Evaluation Methodology

3.1 Sample and scheme

We evaluate out–of–sample forecasts over the long period 1977:Q1 to 2001:Q4 and several subdivisions of that period. The first out–of–sample forecast at horizon τ is for the period ending in 1977:Q1 + τ - 1. After allowing for initial observations set aside for lag determination and differencing, the first forecast is based on models estimated with data from 1958:Q3 + τ - 1 through 1976:Q4. Therefore, in the notation used in Clark and McCracken (2001, 2002a,b), the number of "in–sample" and "out–of–sample" observations are, respectively, $R = 74$ and $P = 100$.[10] Because prior studies have considered a variety of forecast

[8] Under this benchmark, the "random walk" forecast of $\pi_{t+\tau}^{(\tau)}$ is just $\pi_t^{(\tau)}$. Translated into the changes in inflation that are the predictands in our analysis, the random walk–based forecast of $\pi_{t+\tau}^{(\tau)} - \pi_t$ is $\pi_t^{(\tau)} - \pi_t$. In the $\tau = 1$ case, of course, the random walk forecast of the relevant change is 0.

[9] For the core PCE price index, we use just one lag of the change in inflation and one lag of the output gap.

[10] In the case of the core PCE price index, $R, P = 66, 100$.

samples and because subsample analysis can help with identification of structural breaks, we divide the full forecast period into subsamples of: 1977-89, 1990-2001; and 1977-84, 1975-84, 1985-92, 1993-2001.

In our reported results, out–of–sample forecasts from models (2) and (1) are generated recursively. Under this recursive scheme, the parameters of the models are reestimated with added data as forecasting moves forward through time.[11] From these models, we obtain two sequences of $P-\tau+1$ forecast errors, denoted $\hat{u}_{1,t+\tau}$ and $\hat{u}_{2,t+\tau}$, respectively. In results not reported in the interest of brevity, we found that a rolling forecast scheme — estimating the model parameters with a rolling window of the most recent R observations — yields similar results.[12]

3.2 Tests

We base our forecast evaluation on simple RMSE comparisons and four formal tests, including two tests of equal forecast accuracy and two tests for forecast encompassing. In particular, we consider the t–statistic for equal MSE developed by Diebold and Mariano (1995) and West (1996) and the F–statistic proposed by McCracken (2000). We also consider the t–statistic for encompassing developed in Harvey, Leybourne, and Newbold (1998) and West (2001) and the F–type encompassing test proposed by Clark and McCracken (2001). In discussing the results, we focus on the F–type tests proposed by McCracken (2000) and Clark and McCracken (2001) (denoted MSE-F and ENC-F below), as Clark and McCracken (2001,2002a,b) have shown them to be more powerful than their t–type counterparts. To provide something of a benchmark for the forecast tests, we provide results for a standard, full–sample test of causality — the t–statistic on the output gap in (1).

As detailed in the original sources and Clark and McCracken (2001a), the t–statistic for equal MSE takes the form

$$\text{MSE-t} = (P-\tau+1)^{1/2} \times \frac{\bar{d}}{\sqrt{\hat{S}_{dd}}}, \tag{3}$$

where $\hat{d}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}$, $\bar{d} = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\hat{d}_{t+\tau} = MSE_1 - MSE_2$, $\hat{\Gamma}_{dd}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T-\tau}(\hat{d}_{t+\tau}-\bar{d})(\hat{d}_{t+\tau-j}-\bar{d})$, and $\hat{S}_{dd} = \hat{\Gamma}_{dd}(0) + 2\sum_{j=1}^{\bar{j}}(1-\frac{j}{\bar{j}+1})\hat{\Gamma}_{dd}(j)$. As defined here, we compute the long–run covariance $\hat{S}_{dd}$ using the Newey and West (1987)

[11] Formally, for $t = R, \ldots, T-\tau$, forecasts of $\pi^{(\tau)}_{t+\tau} - \pi_t$ use parameter estimates based on data from the beginning of the sample through period t.

[12] With the rolling scheme, allowing the lag orders to change at each point in the forecast sample can introduce additional volatility into forecast performance.

HAC estimator with a bandwidth of $\bar{j} = 1.5 * \tau$, except that in the case of $\tau = 1$, we impose $\bar{j} = 0$. The F–type test of equal MSE takes the form

$$\text{MSE-F} = (P - \tau + 1) \times \frac{MSE_1 - MSE_2}{MSE_2} = (P - \tau + 1) \times \frac{\bar{d}}{MSE_2}. \tag{4}$$

The t–statistic for forecast encompassing — a t–statistic for the covariance between $u_{1,t+\tau}$ and $u_{1,t+\tau} - u_{2,t+\tau}$ — is computed as

$$\text{ENC-t} = (P - \tau + 1)^{1/2} \times \frac{\bar{c}}{\sqrt{\hat{S}_{cc}}}, \tag{5}$$

where $\hat{c}_{t+\tau} = \hat{u}_{1,t+\tau}(\hat{u}_{1,t+\tau} - \hat{u}_{2,t+\tau})$, $\bar{c} = (P - \tau + 1)^{-1} \sum_{t=R}^{T-\tau} \hat{c}_{t+\tau}$, $\hat{\Gamma}_{cc}(j) = (P - \tau + 1)^{-1} \sum_{t=R+j}^{T-\tau} (\hat{c}_{t+\tau} - \bar{c})(\hat{c}_{t+\tau-j} - \bar{c})$, and $\hat{S}_{cc} = \hat{\Gamma}_{cc}(0) + 2 \sum_{j=1}^{\bar{j}} (1 - \frac{j}{j+1}) \hat{\Gamma}_{cc}(j)$. In this case, too, long–run covariance $\hat{S}_{cc}$ in the denominator is the Newey and West (1987) estimator. The encompassing test developed by Clark and McCracken (2001) takes the form

$$\text{ENC-F} = (P - \tau + 1) \times \frac{\bar{c}}{MSE_2}. \tag{6}$$

3.3 Inference

We compare the MSE-F, MSE-t, ENC-F, and ENC-t statistics against critical values generated with the simple parametric bootstrap described in Clark and McCracken (2002a).[13] McCracken (2000) and Clark and McCracken (2001) show that, for one–step ahead forecasts, under the null that the restrictions imposed on model 2 relative to model 1 are true, the tests generally have non–standard, pivotal asymptotic distributions.[14] In particular, the tests converge in distribution to a function of stochastic integrals of quadratics of Brownian motion. Clark and McCracken (2002a) show that, for multi–step forecasts from long–horizon regressions like those in (2) and (1), the tests generally converge to similar, non–standard null distributions, but that the distributions are not pivotal. Clark and McCracken (2002a) go on to show that the simple bootstrap yields good size and power properties for a range of realistic DGPs. We have also verified that, for data–generating processes based on the inflation and output gap models considered in this paper, the bootstrap approach yields tests with good size properties.

Our bootstrap algorithm consists of sampling from the residuals of restricted vector autoregressive equations for the one–quarter change in inflation ($\Delta\pi_t$) and the output gap

[13] The bootstrap is similar to those in Mark (1995) and Kilian (1999).

[14] Clark and McCracken (2002b) and Inoue and Kilian (2002) work out the limiting distributions under certain alternative hypotheses.

(y_t). The inflation equation of this bootstrap DGP imposes the null that the output gap has no predictive power for inflation, taking the form of the restricted model (2) with $\tau = 1$. The output gap equation also takes the form of a simple AR model; the reason is that, allowing different lag orders for the lagged change in inflation and the output gap, the SIC is minimized for 0 lags of inflation. The DGP equations are estimated by OLS using the full sample of data. The initial observations — those preceding the sample of data used to estimate the models — necessitated by the lag structures of the estimated models are selected by sampling from the actual data. In particular, following Stine (1987), the initial observations are selected by picking one date at random and then taking the necessary number of observations in order from that date backward. Each bootstrap replication concludes with forming artificial forecasts using the models and forecasting scheme described above and then constructing artificial test statistics. Following Kilian (1999) and Clark and McCracken (2002a), the number of bootstrap draws is 2000.

For the standard causality test and all of the forecast–based tests, we follow recent conventions and conduct one–sided tests. In our nested model environment, if the restrictions imposed on (2) are true, there is no reason to expect forecasts from that model to be more accurate than those from (1). As a result, for the equal accuracy tests, the alternative hypothesis is that $MSE_1 > MSE_2$, making the test one-sided to the right; the null hypothesis is $MSE_1 \leq MSE_2$. For the encompassing tests, Harvey, Leybourne, and Newbold (1998) point out that, under the alternative that the output gap contains added information, the covariance in the numerator of the encompassing tests should be positive. Under the null that (2) forecast encompasses (1), the covariance of interest will be less than or equal to 0. So the encompassing tests are also one–sided, to the right. Finally, as with the forecast tests, the in–sample t-statistic for causality is one–sided, because the sensible alternative hypothesis is that the gap coefficient should be positive. Inoue and Kilian (2002) advocate the use of one–sided tests because of their power advantage over two–sided tests.

4 Sample Forecasting Results

Using the metrics described above, we obtain sample results consistent with a broad reading of the literature cited in the introduction: on an in–sample basis, the output gap has highly significant explanatory power for inflation, but out–of–sample, the predictive power of the output gap is intermittent. As shown in Table 1's full–sample estimates of (1), the coefficient

on the output gap is highly significant, for both measures of inflation and both measures of the output gap, at both the one and four quarter forecast horizons.[15] Because the CBO–based gap measure is more volatile than the HP–filtered series, the gap coefficient is considerably larger for the HP filtered series, while the $\bar{R}^2$'s are little different.

Despite having statistically strong in–sample explanatory power, on an out–of–sample basis the predictive content of the output gap is prone to breakdowns over certain samples. While the Phillips Curve/AR model RMSE ratios are less than 1 and the test p–values are below 10% in many of the entries in Tables 2 and 3, it is also frequently the case that the RMSEs and p–values exceed 1 and 10%, respectively. For example, the left half of Table 2 shows that, with the GDP price index measure of inflation and the CBO measure of the output gap, the gap has forecasting power until the 1990s. With a forecast horizon of $\tau = 1$, the RMSE ratio for the Phillips Curve relative to the AR model is .973 for 1977-84 and .947 for 1985-92 but .995 for 1993-01; with a horizon of $\tau = 4$, the RMSE ratio rises from .897 over 1977-84 and .841 over 1985-92 to 1.070 over 1993-01. Similarly, the MSE-F test has a p–value of 3 percent or less for all samples but 1993-01, the one sample for which the test fails to reject the null of equal forecast accuracy. The ENC-F test, on the other hand, rejects for all samples. To take another example, the right half of Table 3 shows that, with core CPI inflation and the HP–filter measure of the output gap, the gap's predictive content breaks down sharply in the mid-1980s and, to a lesser extent, the 1990s. At the one–quarter horizon, the RMSE ratio is .951 for 1977-84 but 1.025 for 1985-92 and 1.009 for 1993-01, and the MSE-F test rejects the null of equal MSE for all samples shown except 1985-92 and 1993-01. At the four–quarter horizon, the RMSE ratio rises from .978 over 1977-84 to 1.230 over 1985-92 and declines to .925 over 1993-01. The MSE-F test fails to reject the null for samples except 1990-01 and 1993-01.

Tables 2 and 3 yield two other notable results. First, even when the output gap has significant out–of–sample explanatory power, the forecast gains are quantitatively modest. For example, over the full forecast sample of 1977-2001, including the CBO output gap in the models for forecasting inflation in the GDP price index and core CPI yields a roughly 3% reduction in RMSE at the one–quarter horizon. At the four-quarter horizon, including the CBO output gap in the model lowers the RMSE about 10% for inflation in the GDP

[15] As there is some concern with the small–sample properties of coefficients in long–horizon regressions, we also used our bootstrap to evaluate the significance of the coefficients, and consistently obtained p-values of 1 percent or less.

price index and 5% for core CPI inflation. Second, the ENC-F test of forecast encompassing rejects the null of no predictive content in the gap much more frequently than the MSE-F test does. Of the 48 tests reported in Tables 2 and 3 (these observations are of course not independent), there are 16 instances in which the ENC-F test rejects the null at 10% confidence but the MSE-F does not. For example, in the case of the GDP price index and HP–filtered gap, at the four–quarter horizon the MSE-F test rejects the null for only the 1990-01 and 1993-01 samples, while the ENC-F test rejects the null for all six samples.

5 Evaluating the Power Explanation

Especially in light of the evidence that the output gap only modestly improves the accuracy of inflation forecasts, one possible explanation for the weakness of our out–of–sample evidence relative to the in–sample evidence is *power*. It could simply be that the out–of–sample evidence is relatively weak because the out–of–sample tests are less powerful than the usual in–sample causality test. Inoue and Kilian (2002) advocate the use of in–sample tests for model evaluation because of their power advantages.

In this section we use simulation methods to evaluate the extent to which power differences can account for the patterns in our sample results. In particular, we specify data–generating processes based on models fit to our full sample of data, in which the gap has significant (in–sample) explanatory power for inflation, and conduct Monte Carlo simulations to evaluate the power of our forecast metrics. The inflation equations in the DGPs, which impose stability in the DGP, are those given in Table 1 for $\tau = 1$. The gap equation in each DGP takes a simple autoregressive form; in the sample estimates, allowing separate lag lengths on the change in inflation and the gap in the equation, the SIC always selects 0 lags of inflation for the gap equation. Artificial data for 1958:Q3–2001:Q4 are generated by drawing innovations from the standard normal distribution, with a variance–covariance matrix equal to that estimated from the sample data. The initial observations of the artificial data on $\Delta\pi$ and y are taken from draws of the unconditional normal distribution implied by the parameterization of the DGP.

In each Monte Carlo draw, we first form artificial forecasts using models (1) and (2) and compute the test statistics of interest. For each artificial data set, we then apply our bootstrap method and compute 10% bootstrap critical values against which the "sample" test statistics from the artificial data set are compared. Power is then measured as the

percentage of Monte Carlo simulations in which the artificial "sample" test statistics exceed their bootstrapped critical values.

In broad terms, the simulation results in Tables 4 and 5 show that while some of the gap in our sample evidence on in–sample and out–of–sample predictive content can be attributed to the lower power of out–of–sample metrics, much of the gulf cannot be explained by power. Omitted from the tables for simplicity, the simulated power of the in–sample t–statistic is at least 97 percent for all of the DGPs and forecast horizons covered in Tables 4 and 5. While the overall power of the forecast–based tests seems quite good, in some instances the powers of the out–of–sample tests over subsamples fall well below that of the in–sample test. In many cases, the power of the out–of–sample tests comes near the power of the in–sample t–test. For example, in the DGP based on the GDP price index and CBO output gap, for 1–step ahead forecasts the MSE-F test has power of 90.8% over the full forecast sample of 1977-01 and 79.2% over the subsample of 1990-01 (Table 4, row 1, left panel). But there are also instances of shortfalls. In the same DGP, the MSE-F test has power of 66.7% for 1–step forecasts and 63.5% for 4–step forecasts over 1977-84. The corresponding figures fall to 62.5% and 54.8% when the core CPI replaces the GDP price index (Table 5, row 1). Consistent with the results in Clark and McCracken (2001,2002a), the power of the ENC-F test exceeds that of the MSE-F test. The differences are small in some cases and large in others.[16] Accordingly, the sample finding that the ENC-F test rejects more often than the MSE-F test does is partly a function of an underlying difference in the powers of the tests.

The results in Tables 4 and 5 confirm the relationships of power to data sample sizes and to forecast horizon described in Clark and McCracken (2001,2002a). The powers of the tests generally fall as the forecast horizon rises from $\tau = 1$ to $\tau = 4$. The powers of the tests also rise with the number of forecast observations. For example, rejection rates are greater for the forecast sample of 1977-01 than for 1977-89, and for 1977-89 than 1977-84. Given the number of out–of–sample observations, power also tends to rise with the number of in–sample observations (denoted R in our previous work) used to estimate the forecasting model. For instance, the powers of the tests are slightly higher for the forecast subsample of 1993-01 than for 1985-92.

The power properties of the out–of–sample tests — especially MSE-F — are reflected

[16]Consider, for example, the DGP based on the GDP price index and HP gap and a forecast horizon of $\tau = 1$. With a forecast sample of 1977-01, the MSE-F and ENC-F powers are 95.1% and 98.8%, respectively. But with a sample of 1993-01, the MSE-F and ENC-F powers are 76.3% and 93.3%.

in the summary statistics for the RMSE ratio distribution reported in Tables 4 and 5. On average, including the output gap in the forecasting model only modestly lowers the forecast RMSE. The gains in accuracy indicated by the median RMSE ratios range from 3% to 5% for $\tau = 1$ and 7% to 10% for $\tau = 4$. But reflecting the generally good power of the equal MSE and encompassing tests, the probability of the RMSE ratio being less than 1 is reasonably high. The probability is highest for the full 1977-01 forecast sample, ranging from 87% to 96% (including both forecast horizons), but still decent over the subsamples, falling no lower than 70%. Nonetheless, the 90% confidence bands for the RMSE ratios are wide enough that they almost always include a value of 1. For example, as reported in Table 5 for a DGP based on the core CPI and CBO output gap, the 90% confidence band for the 1977-01 RMSE ratio is (.912,1.010) for 1–step ahead forecasts, with a median of .970. The 90% bands are considerably wider for 4–step ahead RMSEs, with most of the widening coming in the left tail. For the same DGP and sample period, the 90% confidence band for the 4–step ahead RMSE ratio is (.788,1.033), with a median of .919.

Finally, a detailed comparison of the simulation results in Tables 4 and 5 with the sample results in Tables 2 and 3 shows that a stable DGP can only account for a portion of the sample results. In the case of the GDP price index and CBO output gap, the sample results seem at least partly consistent with the stable DGP simulations. The sample RMSE ratios all fall within the simulated 90% bands. Nonetheless, the simulations show that a stable DGP is unlikely to produce a breakdown in the 1990s: there is no falloff in simulated power from 1977-84 or 1985-92 to 1993-01, and the MSE-F test has power of more than 72 percent in 1993-01. When inflation is measured with the core CPI, the stable DGP has similar difficulty in explaining a similar sample breakdown in predictive power in the 1990s. In this case, though, the shortcomings of the stable DGP are greater: the sample values of the RMSE ratios for 1993-01 greatly exceed the upper ranges of the simulated 90% bands.

With the HP–filtered measure of the output gap, the sample results are clearly at odds with the simulation results. For the GDP price index and HP–filtered output gap combination, the sample RMSE ratios exceed the upper range of the simulated 90% bands in five instances: for 1–step forecasts over 1977-01, 1977-89, and 1977-84 and for 4–step forecasts over 1977-01 and 1977-89. Moreover, while the data show that the MSE-F test overwhelmingly fails to reject the null for 1977-01, the stable DGP simulations indicate the test has power of 90 percent or more. When inflation is measured with the core CPI, there is one

instance in which the sample RMSE ratio exceeds the simulated 90% band: 4–step ahead forecasts over 1993-01. In this case, too, the simulated powers of the tests seem too high to be consistent with the breakdowns observed in the sample results. For example, in the data, the MSE-F test for $\tau = 1$ rejects the null over 1977-84 but not 1985-92 or 1993-01. But the simulated power of the test rises from 72.3% for 1977-84 to 75.9% for 1985-92 and 77.6% for 1993-01.

6 Evidence of Structural Breaks

With in–sample vs. out–of–sample power differences in stable DGPs explaining only a portion of our empirical results on the predictive content of the output gap for inflation, we turn now to an alternative explanation: discrete breaks in the parameters of the Phillips curve models. On the surface, at least, the rolling estimates of output gap coefficients shown in Figure 1 suggest instabilities in the gap coefficient (the figure reports the gap coefficient from estimates of (1) for $\tau = 1$ based on a rolling window of 74 observations). Clark and McCracken (2002) present theoretical and Monte Carlo evidence on how breaks in the causal coefficients of interest can lead to out–of–sample breakdowns in predictive content, with examples that use financial variables to predict output growth. Here, we more generally consider the effects of shifts in not only the causal coefficients of interest (those on the output gap) but also shifts in the lagged dependent variable coefficients (those on lagged inflation) and the residual variance. And, in this case, we take a much more detailed look at a different empirical relationship, the Phillips Curve.

We limit our attention in the discussion below to discrete breaks rather than stochastic time variation (TVP) in the parameters for several reasons. First, in unreported results, we found little evidence of TVP in the full–sample estimates of our Phillips curve. Nyblom (1989) tests for the null of stability against the alternative of unit root variation yielded virtually no rejections of the null. Second, auxiliary simulations provided little evidence that TVP can account for the sample results. Simulations of models in which the output gap coefficient follows a random walk show that TVP is unlikely to generate the breakdowns in out–of–sample predictive content found in the data over certain subsamples.[17] Finally, it is worth noting that, in line with the evidence in Stock and Watson (1996), allowing TVP

[17]In these experiments, we impose the residual variance and discrete AR coefficient breaks described below. The output gap coefficient follows a random walk process with initial value equal to the full–sample OLS estimate of the gap coefficient and innovation variance equal to that of the residual in a regression of the rolling sample estimate of the gap coefficient on one lag of itself.

failed to yield more accurate sample forecasts of inflation.

In the results reported in this section, we test for breaks in the coefficients and residual variances of the models under consideration using Andrews' (1993) extremum Wald test and Bai and Perron's (1998, 2002) tests for multiple breaks. In light of the strong evidence in Stock and Watson (2002) of shifts in the volatility of inflation, we follow Hansen (2001) and Stock and Watson (2002) and use heteroskedasticity–robust variances in forming Andrews' Wald statistics for shifts in the regression coefficients. Break tests are computed for individual coefficients and the set of lagged inflation coefficients in each equation, as well as the residual variance. We compare the test statistics for a single break against Andrews' asymptotic critical values and report p-values computed with Hansen's (1997) asymptotic approximation.[18] In investigating multiple breaks, we follow the recommendations of Bai and Perron (2002) and use their $UD\max$ and $WD\max$ tests to determine whether there are any significant breaks and their sup $F(l+1|l)$ statistics to determine the number. In reporting Bai–Perron test results, we simply provide the test statistics and significance levels indicated by the Bai-Perron critical values. In all cases, for simplicity we apply the breaks tests to just the models for 1-step ahead forecasting.

Overall, our break test analysis yields stronger evidence of shifts in the autoregressive coefficients of the Phillips curve than of shifts in the coefficient on the output gap, in line with the results of Stock and Watson (1999b, 2001). Given the imprecision in break dating, in this section we include some additional evidence, for the core PCE price index. As shown in Table 6's Andrews test results, the four models using either the core CPI or core PCE measures of inflation have significant breaks in the set of coefficients on lagged inflation. The break in the CPI coefficients is estimated to occur in 1980, while the break in the PCE coefficients is estimated in 1983. It should be noted, though, that these breaks are not statistically significant based on p-values computed with Hansen's (2000) fixed regressor (heteroskedastic) bootstrap. While Andrews' (1993) asymptotic critical values remain the default in studies conducting breaks testing, Hansen's bootstrap offers the advantage of generally and appropriately accounting for the presence of lagged dependent variables as regressors. Even when gauged against asymptotic critical values, the two models using the GDP price index measure of inflation do not suffer a significant break in the set of AR coefficients. In results not reported in the interest of brevity, Bai–Perron tests show that

[18] Generating p-values with Hansen's (2000) fixed regressor (heteroskedastic) bootstrap yields generally similar results, with one exception noted in the discussion below.

the number of breaks in the autoregressive coefficients is at most 1. Collectively, there appears to be modest evidence of a single break in the autoregressive coefficients of our Phillips curve formulations.

The results in Table 6 show that formal evidence of breaks in the output gap coefficients is even weaker. With inflation measured using the GDP or core PCE price index, the p–values of the Andrews test range from .27 to .73. In some unreported estimates, we also found that redoing the tests imposing a single break in the AR coefficients in the early 1980s does not alter this result. In some other unreported results for these models, applying the Bai–Perron tests failed to yield any evidence of multiple breaks. The only formal evidence of instability in the output gap coefficient is obtained with the core CPI measure of inflation. As shown in Table 6, the Andrews test identifies a significant break in the output gap coefficient in 1983, for both the CBO and HP estimates of the gap. In the CBO case, Bai–Perron tests even provide some evidence of two significant breaks in the gap coefficient, with one break in 1983 and the other in the early 1970s.

We find much stronger evidence of breaks — indeed, multiple breaks — in the residual variance of the Phillips curve.[19] As shown in Table 7, Bai–Perron tests provide strong evidence of at least one break in the residual variance of the models. In these reported results, we have imposed the regression coefficient break specifications described below in section 7.2, but the results are insensitive to the specification of breaks in the regression coefficients. For example, testing for breaks in the residuals of stable AR models for inflation yields very similar results. As shown in the table, the UDmax and WDmax tests are all highly significant. The $supF(l+1|l)$ tests indicate there are three breaks in the variances of the core CPI equations, with estimated dates of 1974:1, 1984:1, and 1991:4.[20] These dates seem to capture the low–frequency movements in inflation volatility suggested by the squared residuals plotted in the upper panel of Figure 2. For the GDP price index equations, the evidence on the number of breaks is a bit less clear, but overall there appear to be two breaks. Again, highly significant UDmax and WDmax tests indicate the presence of some breaks. As to the number, the SupF(2|1) statistic is significant at the 10% level when the CBO–based output gap is used and not quite significant at 10% when the HP–filtered gap

[19]These tests use the HAC correction option built into the Bai–Perron program. Similar results are obtained when the residual variance of the break test equation — which corresponds to the centered fourth moment of the inflation equation residual — is restricted to be the same across break segments.

[20]These dates globally minimize the sum of squared errors. The dates are the same for the CBO and HP measures of the gap, except that the HP–based gap specification puts the middle break a quarter earlier, at 1983:4.

is used. While the SupF(1) test is not at all significant, the SupF(2) statistic rejects the null of no breaks against the alternative of two breaks at the 1% level. The estimated dates of the two breaks are 1970:2 and 1976:2. As shown in the lower panel of Figure 2, these break dates capture the elevated volatility of the early 1970s.

7 Evaluating the Structural Breaks Explanation

In this section we use the break test results described above in evaluating whether structural breaks can account for the in–sample and out–of–sample results presented in section 4. With the evidence of breaks strongest for the residual variances, a natural starting point is to consider models with stable coefficients but residual variances having the breaks identified above. But it turns out that such breaks, by themselves, do little to improve upon the stable DGP explanation for our sample results. In results not reported in the interest of brevity, we conducted simulations like those described in section 5, with DGPs having the same, stable coefficients, but residual variances with breaks. These simulations show that, when inflation is measured with the GDP price index, the variance breaks might be seen as helping a little to explain the sample results, but not much. Introducing breaks in residual variances into the DGP serves to lower the power of the forecast tests over both 1985-92 and 1993-01 relative to 1977-84. Such changes help, but still fall short of accounting for the sample results. For example, the simulations fail to show the kind of distinct falloff in predictive power over 1993-01 seen in the sample results for the CBO gap case. When inflation is measured with the core CPI, the variance breaks widen the gap between the sample and simulation results. For example, introducing the breaks causes the power of the forecast–based tests to rise sharply from 1977-84 to 1985-92 and in turn to 1993-01.

7.1 Breaks in just AR coefficients

With breaks in residual variances unable to generate the patterns seen in the sample results, in this section we consider whether additional breaks in the autoregressive coefficients of the Phillips curve can account for the sample results. We report the results of simulations of DGPs that allow for the identified breaks in the residual variances and identified or imposed breaks in the autoregressive coefficients of the inflation equation. Our simulation methodology is the same as described in section 5, except that the DGPs now incorporate shifts in the AR coefficients, with the shifts corresponding to those estimated in the data.

In simulating models with breaks in the coefficients on lagged inflation, we use empirically–identified dates and model estimates, (the latter) reported in Table 8. In particular, for the core CPI, we use a break date of 1980:2, as identified by the significant (depending on the critical values used) break tests discussed above. For the GDP price index, we simply impose a break date of 1983:1, even though the formal tests fail to identify a significant break. Our rationale for imposing the break is three-fold. First, break tests for the core PCE price index, which represents more than half of the GDP price index, indicate a significant (again, depending on critical values) break in 1983. Second, point estimates allowing a 1983 break in the GDP price index coefficients suggest a quantitatively large, albeit insignificant, shift. Finally, our simulations of models with breaks in the AR coefficients indicate that the finite–sample power of the break test is at best modest. For example, in the GDP price index and CBO gap case, simulations of the DGP with the 1983 break in the AR coefficients show that the Andrews test for a break in the set of AR coefficients has power of just 46% (when the test is compared against the 10% asymptotic critical value). Note that, in the AR model for the output gap used in the simulated DGPs, we treat the coefficients as stable and allow a single break in the residual variance in 1984:1. Tests for the stability of the AR coefficients yield mixed results, with some evidence of stability and some evidence of a break. With the evidence mixed, we treat the gap equation coefficients as stable for simplicity. Tests for the stability of the residual variance of the gap equation provide strong evidence of a break, in 1982 or 1984 depending on the gap measure and test particulars. With the core CPI having an variance break in 1984, for simplicity we impose a break date of 1984:1 on the output gap equation's residual variance.

The simulation results reported in Tables 9 and 10 show that breaks in the AR coefficients and residual variances cannot account for the sample results. While introducing the breaks in the AR coefficients helps along some dimensions, it hurts along others. With inflation measured using the GDP price index, introducing the breaks helps in the sense that power is flat or declines a bit from 1977-84 to 1985-92 to 1993-01, while the stable DGP simulations show power rising across subsamples. Nonetheless, the AR and variance breaks fall far short of generating the kind of 1993-01 breakdown in forecast test power seen so dramatically in the sample results for the CBO output gap. Moreover, with the HP–filtered output gap, introducing the AR breaks fails to account for the large RMSE ratios in the sample data including the late 1970s and early-to-mid 1980s — a number of

the sample RMSE ratios significantly exceed the upper range of the simulated 90% band. With inflation measured by the core CPI, introducing the breaks helps account for some of the RMSE ratios that exceed 1 in the sample, but fails to account for the sample pattern of test results. As shown in Table 10, the simulated powers of the tests rise significantly from 1977-84 to 1985-92 to 1993-01. Yet, in the data, the predictive content of the CBO output gap breaks down in 1993-01, while the predictive power of the HP–filtered gap suffers breakdowns in both 1985-92 and 1993-01.

7.2 Breaks in both AR and output gap coefficients

With breaks in the lagged inflation coefficients and residual variances helping little, and with the results in Clark and McCracken (2002b) highlighting the potential effects of breaks in causal coefficients, we now turn to examining whether breaks in the output gap coefficients can account for our sample results. As noted in the discussion of the break test results in section 6, there is only weak direct evidence of breaks in the gap coefficients, with the only rejections of stability occurring for the core CPI equations. We nonetheless pursue the gap coefficient break route because (1) such breaks seem to be a potentially good explanation for the sample forecasting results and (2) our simulations show that the finite–sample power of tests for a break in the gap coefficient is often modest, a point we return to below. In this section we report the results of simulations of DGPs that allow for discrete breaks in not only the residual variances and autoregressive coefficients but also the output gap coefficient.

In light of the weak test results, often modest power of the break test, and well–known imprecision in estimates of break dates, we have cast a wide net in trying to come up with plausible breaks explanations for the sample results. We generally started out with break dates in accord with the dates at which the standard test statistics are maximized. We considered both single breaks identified by Andrews tests and dual breaks identified with the Bai-Perron methodology. In some cases, we also tried break dates suggested by results for another model — for example, imposing on the GDP price index a break date actually identified for the core PCE price index (again, the core PCE price index amounts to about half of the GDP price index). Our choices have also been guided by a preference for a simple and consistent set of breaks or explanations. For instance, in some cases, while a two–break specification could explain the sample results nearly as well as a one–break specification, we favor the one–break specification for its simplicity. As a particular example, for the core

CPI and CBO output gap model, Bai–Perron tests provide some evidence of two breaks in the gap coefficient, which we date as 1973 and 1983. But we find that a single break in 1994 explains the sample results for this inflation-gap combination somewhat better and also provides an explanation for the GDP price index–CBO gap and core PCE price index–CBO gap results. For simplicity, we report results for just those break specifications that best account for the sample patterns and explain in the discussion how we arrived at the break specifications.

In light of this approach, we must stress that our results should simply be viewed as general evidence that discrete breaks in the output gap coefficient can account for the sample results, rather than as definitive explanations for the sample results. Breaks identification is fraught with small sample problems (Diebold and Chen (1996) document size and power problems in testing; a number of authors have noted the imprecision in dates estimates). Moreover, as this discussion indicates, we have engaged in a considerable specification search in trying to account for the sample results, trying a range of possible break dates and reporting, in effect, the "best" results.

7.3 Break specifications

Table 11 reports Phillips Curve estimates that impose the break dates that prove to best explain our sample results on the predictive content of the output gap. The models also incorporate breaks in the AR coefficients at the dates identified in the prior subsection, 1983:1 in the case of the GDP price index and 1980:2 in the case of the core CPI. For models that use the CBO output gap, the estimates in the table impose a single break in the gap coefficient in 1994:1. We initially considered this date for the core PCE price index, because it corresponded to the SupWald test date, at which the regression estimates indicated a quantitatively large coefficient shift (despite the insignificance of the SupWald statistic). In the case of the GDP price index, Table 11's regression estimates show that the gap coefficient declines only slightly (and insignificantly) over 1994-01, but is very imprecisely estimated. To account for the sample breakdown in predictive power in the 1990s, we need to impose a larger coefficient break. With the standard error so large, we simply use a DGP with a gap coefficient of 0, corresponding to an estimated regression with the coefficient restricted to 0 over 1994-01 (imposing the restriction leaves the other coefficients hardly changed from those reported in the table). A DGP with a coefficient restricted to a small, positive value such as .04 over 1994-01 yields very similar simulation

results. In the case of the core CPI, the regression estimates show a much larger, more precisely estimated decline in the CBO gap coefficient.[21] For consistency, we restrict the 1994-01 gap coefficient to 0 in the equation used in the simulated core CPI–CBO gap DGP.

For models that use the HP–filtered output gap, the estimates in Table 11 impose a single break in the gap coefficient corresponding to the date associated with the SupWald test. In particular, the estimates for the GDP price index impose a break in 1976:2, producing a fall in the gap coefficient from .328 over 1958-76 to .119 over 1976-2001. The estimates for the core CPI impose a break in 1983:1, one which the Andrews test results in Table 6 identified as statistically significant (with asymptotic critical values). The estimated gap coefficient falls from .359 before the break to essentially 0 after the break, with a large standard error. But a reduction in the coefficient to 0 turns out to be inconsistent with the sample results in which the gap retains some predictive power for inflation in the late 1980s and 1990s. Given that, and the imprecision of the regression estimate of the post 1983 coefficient, in our simulations we impose a smaller break in the HP gap coefficient. In particular, in the simulated DGP we restrict the 1983-01 coefficient to have a value of .10, with virtually no effect on the restricted estimates of the other coefficients of the model.

Notwithstanding the considerable uncertainty surrounding the specification of break dates, simulations of these DGPs yield break test results that suggest our break date configurations to be reasonable. Consistent with the sample evidence that showed no significant break in the gap coefficient in the GDP price index–CBO gap case, the break test appears to have little power to identify the imposed break in 1994. Using 10% asymptotic critical values, the simulated rejection rate is only 26%. But in the GDP price index–HP gap case, the simulated power of the test is much higher — 73%, with an average date estimate of 1975:4.[22] On the one hand, this rate is high enough that, if there were such a break, it should have been evident in the test in the data. On the other hand, while the sample test fails to indicate a significant break, the estimated regression shows a large coefficient change. For the core CPI–based models, the simulated behavior of the break test seems in line with the sample evidence. In the data, the Andrews test indicates a significant break in the early 1980s. In the simulations, a break test applied to the gap coefficient rejects the null of stability with a frequency of 61% in the case of the CBO gap and 90% in the

[21] In a result not shown in the table, the standard error on the -.120 change in the coefficient is .062.

[22] The average date estimate is a median of the distribution of supWald tests generated in the 1000 Monte Carlo simulations.

case of the HP gap, with average date estimates in the early 1980s (the match between the simulated average date and the sample estimate date is an especially good outcome in the case of the CBO gap, as the DGP in this case imposes a break in 1994).

7.4 Simulation evidence on the breaks explanation

Though not necessarily definitive explanations of the sample results presented in section 4, the breaks we have identified or imposed on the output gap coefficients reasonably account for the sample results. As shown in the simulation results reported in Tables 12 and 13, the sample RMSE ratios all lay within the simulated 90% bands (in contrast, in the stable DGP case, three of the four models had sample RMSEs outside the simulated bands). This finding implies that each of the sample RMSE ratios is consistent with the distribution the ratio would have if the true DGP had the imposed break in the output gap coefficient. Moreover, in the two models using the CBO measure of the output gap, the break in the gap coefficient in 1994 induces a significant falloff in the power of the forecast–based tests for 1993-01, more so for MSE-F than ENC-F, in line with the sample finding that ENC-F rejects but MSE-F does not in the 1993-01 period. For example, as shown in Table 12, with inflation measured by the GDP price index, the simulated power of the MSE-F test for 1–step ahead forecasts declines from 79.7% in 1977-84 and 76.0% in 1985-92 to 46.9% in 1993-01, while the power of the ENC-F test declines by 91.7% to 86.8% to 69.1% over the same periods. Table 13 shows that, with inflation measured by the core CPI, power rises from 1977-84 to 1985-92 and then falls sharply in 1993-01; the rise in power from 1977-84 to 1985-92 stems from the sharp reduction in innovation variance in 1984 that the DGP incorporates.

Similarly, in the two models using the HP–filtered measure of the output gap, the imposed breaks in the output gap coefficient appear to largely account for the sample results. As reported in Table 12, when inflation is measured by the GDP price index, the simulated power of the MSE-F test is significantly higher for 1990-01 (66.6%) than 1977-89 (43.6%) and much higher for 1985-92 (64.9%) and 1993-01 (62.2%) than 1977-84 (35.5%), in line with the sample results for 1–step forecasts that show the MSE-F rejecting the null for only 1990-01, 1985-92, and 1993-01. And, consistent with the sample results in which the ENC-F test rejects for all samples, the simulated power of ENC-F is always at least 20 percentage points greater than the power of MSE-F. The same basic characterization applies to the results for 4–step ahead forecasts. For core CPI inflation, the simulation results in

Table 13 show the power of MSE-F declining sharply from 1977-84 (77.4% for $\tau = 1$) to 1985-92 (48.0%) and modestly further in 1993-01 (35.0%), while the power of ENC-F is little changed, remaining above 90% across all 1–step ahead subsamples. These simulation results for $\tau = 1$ are consistent with the sample results that show predictive power as gauged by MSE-F, but not ENC-F, falling off in 1985-92 and 1993-01. Admittedly, though, the DGP has a harder time accounting for the 4–step ahead CPI forecasts. In the sample, the MSE-F test rejects the null for only the 1990-01 and 1993-01 periods, but the simulated power of the test is only about 20 percent over these periods. Imposing a gap coefficient slightly larger than the value of .10 we use would boost simulated power over the 1990s period, but would also boost power over the 1985-92 period. Ultimately, in the core CPI–HP gap case, the differences across forecast horizons in the mix of rejections over some subsamples but not others evident in the sample results proves difficult to fully explain.

8 Predictive Ability and Unmodeled Breaks

In this section we consider whether the asymptotic behavior of the tests of predictive ability used above is consistent with our empirical and Monte Carlo evidence. In particular we are interested in whether structural breaks in the residual variance, breaks in the parameters associated with lagged inflation and breaks in the parameters associated with the output gap can affect the power of tests for predictive ability. Since structural breaks are a major component of our argument we also consider whether structural breaks themselves can distort the power of structural break tests. For example it may be that unmodeled structural breaks in the regression parameters distort the power of tests for a break in the residual variance. Similarly, breaks in the residual variance may distort the power of tests for breaks in the regression parameters.

We analyze the power of tests for predictive ability and structural breaks by considering the magnitudes of their noncentrality parameters. We begin with tests for structural breaks, examining the power of a simple Chow test for a single shift in regression coefficients and a simple Goldfeld-Quandt (hereafter, GQ) test for a single change in residual variance. Using these simple tests for a known break point rather than the Andrews (1993) and Bai and Perron (1998, 2002) generalizations used in our empirical and Monte Carlo work greatly simplifies the analysis without affecting the basic conclusions. Similarly, in examining the power of the usual in–sample test for whether the output gap has explanatory power, we

focus on the simple F–test (hereafter referred to as the **GC** test) rather than the one–sided t–test actually used in our results above. Finally, we consider the power of the MSE-F and ENC-F tests for, respectively, equal forecast accuracy and encompassing. For each of these five test statistics we derive the noncentrality parameters implied by a particular sequence of local alternatives and analyze the power of the test through its magnitude. The local alternative allows for the output gap to have predictive content and for the regression parameters and the residual variance to suffer structural breaks.

In this analysis, we extend the results of Clark and McCracken (2002b) in several ways. Clark and McCracken (2002b) consider the limiting distributions of the GC, MSE-F, and ENC-F tests under a general alternative that allows any number of discrete breaks in the coefficient vector. The special case of a single break in the causal coefficient of interest in a rudimentary DGP is used to simplify the results. In this paper, we focus on a sequence of *local* alternatives featuring a single break and disentangle the effects of breaks in the causal coefficient of interest from the effects of breaks in the other coefficients of the model. This paper also extends Clark and McCracken (2002b) by examining whether breaks in one set of coefficients or the residual variance can distort tests for breaks in the coefficients of interest or whether breaks in the coefficients can distort tests for breaks in the residual variance.

Below we introduce some notation and give a broad overview regarding assumptions. In order to focus on the substantive issues of predictive ability we omit the detailed assumptions and restrict attention to models designed for 1-step ahead prediction. Detailed assumptions and proofs can be found in a not-for-publication technical appendix, Clark and McCracken (2003).

8.1 Setup

Suppose we are interested in predicting the covariance stationary scalar y_{t+1} using a $(k_1 + k_2 = k \times 1)$ vector of covariance stationary covariates $x_{2,t} = (x'_{1,t}, x'_{22,t})'$. The problem is that we are unsure whether $x_{22,t}$ is useful for prediction given the conditioning variable $x_{1,t}$. For this paper we have in mind that y_{t+1} is the change in inflation, $x_{1,t}$ is a vector including an intercept and lagged changes in inflation and $x_{22,t}$ is a vector including lags of the output gap. To test this hypothesis it is common to posit that $y_{t+1} = x'_{2,t}\beta_2^* + u_{t+1}$ with $\beta_2^* = (\beta_1^{*\prime}, \beta_{22}^{*\prime})'$ and u_{t+1} a conditionally homoskedastic martingale difference sequence with $Eu_{t+1}^2 = \sigma_u^2$ and $E(u_{t+1}^2 - \sigma_u^2)^2 = V$. A test is then constructed for determining whether $\beta_{22}^* = 0$.

The standard F-test is constructed using the residual sum of squares implied by two linear models, $x'_{i,t}\beta_i$, $i = 1,2$, each estimated by OLS, with model 1 excluding the output gap variables included in model 2. We denote these residuals as $\widehat{v}_{i,s+1} = y_{s+1} - x'_{i,s}\widehat{\beta}_{i,T}$ s = 1,..., T for models 1 and 2, respectively. In contrast, the out-of-sample tests (MSE-F and ENC-F) are constructed using the forecast errors and their squares implied by the same two linear models each estimated by OLS. We denote these forecast errors as $\widehat{u}_{i,t+1} = y_{t+1} - x'_{i,t}\widehat{\beta}_{i,t}$, $t = R, \ldots, T$ for models 1 and 2, respectively.

As mentioned above, in this section we are interested in the behavior of these three tests of predictive ability in the non-standard environment in which there are different types of unmodeled structural change occurring in the background. To analyze this issue we consider the data generating process

$$\begin{aligned} y_{t+1} &= x'_{1,t}(\beta_1^* + T^{-1/2}\Delta\beta_1^*\ 1(t \geq T_B)) \\ &\quad + x'_{22,t}(T^{-1/2}\beta_{22}^* + T^{-1/2}\Delta\beta_{22}^*\ 1(t \geq T_B)) \\ &\quad + (1 + T^{-1/2}\ \delta\ 1(t \geq T_B))u_{t+1} \\ &= x'_{2,t}\beta_{2,T}^* + x'_{2,t}\Delta\beta_{2,t,T}^* + u_{t+1} \\ &= x'_{2,t}\beta_{2,t,T}^* + u_{t+1}, \text{ where } T_B < T \text{ denotes a break point.} \end{aligned}$$

By doing so we capture many non-standard conditions which may interact to distort the power of the tests of predictive ability and for that matter, the tests for structural breaks. These non-standard conditions include breaks in the parameters associated with the conditioning variables ($T^{-1/2}\Delta\beta_1^*\ 1(t \geq T_B)$), breaks in the parameters associated with those variables whose predictive content is in doubt ($T^{-1/2}\Delta\beta_{22}^*\ 1(t \geq T_B)$) and breaks in the residual variance ($T^{-1/2}\ \delta\ 1(t \geq T_B)$). The term that captures the fact that the output gap may have predictive ability is given by $T^{-1/2}\beta_{22}^*$ (and to a lesser extent $T^{-1/2}\Delta\beta_{22}^*\ 1(t \geq T_B)$). Let the location of the break(s) T_B and forecast sample split parameter R satisfy $\lim_{T\to\infty} T_B/T = \lambda_B \in (0,1)$ and $\lim_{T\to\infty} R/T = \lambda \in (0,1)$. For ease of presentation we restrict the timing of the breaks so they occur simultaneously.

Note that, for simplicity, we have omitted the subscript T that is associated with the observable variables of the DGP. For the dependent variable y_{t+1}, the (implicit) subscript T captures the fact that y is being generated using parameter sequences that depend upon the sample size. The predictors $x_{1,t}$ and $x_{22,t}$ have the (implicit) subscript T implicit because we do not want to make the additional assumption that these variables are strictly exogenous to the system. Such an assumption would eliminate the possibility

that lagged changes in inflation were used as predictors and the possibility that lagged changes in inflation were useful for predicting the output gap. Ultimately, the dependence of the observable variables and their marginal distributions on the sample size T is not particularly important when the sample size is large. Because the models are linear and the test statistics are formed using sample averages of quadratics of the residuals and forecast errors, we are able to take advantage of the concept of asymptotic mean square stationarity as discussed in Hansen (2000). For example, when deriving the standard F-test we work with sample averages of quadratics of the observables. Consider the probability limit of $T^{-1}\sum_{s=1}^{T} x_{1,s,T}x'_{1,s,T}$, where the appropriate subscript T made implicit above and below has been added to clarify our point here. Under mild conditions on the marginal distributions of these variables we obtain plim $T^{-1}\sum_{s=1}^{T} x_{1,s,T}x'_{1,s,T}$ $= \lim_{T\to\infty} T^{-1}\sum_{s=1}^{T} E(x_{1,s,T}x'_{1,s,T})$. But for each s, $\lim_{T\to\infty} E(x_{1,s,T}x'_{1,s,T}) = E(x_{1,s}x'_{1,s})$ and hence $\lim_{T\to\infty} T^{-1}\sum_{s=1}^{T} E(x_{1,s,T}x'_{1,s,T}) = \lim_{T\to\infty} T^{-1}\sum_{s=1}^{T} E(x_{1,s}x'_{1,s})$. Since $x_{1,s}$ is covariance stationary we find that plim $T^{-1}\sum_{s=1}^{T} x_{1,s,T}x'_{1,s,T} = E(x_{1,s}x'_{1,s})$ and we obtain the result we would have obtained if the observables did not depend on T.

8.2 Propositions

In the following 5 Propositions we derive the non-centrality parameters for each of the GC, MSE-F, ENC-F, Chow and GQ test statistics. For each of these we use the notation and data generating process described above. Since the possible presence of breaks is the central focus of our discussion we begin with the Chow and GQ tests and then proceed to the tests of predictive ability.

Consider the Chow test for a break in a single regression coefficient from the model $x'_{2,t}\beta_2$. To construct this test we estimate two models using OLS and compare the residual sum of squares in the standard way. The restricted model is $x'_{2,t}\beta_2$ while the unrestricted model is $x'_{3,t}\beta_3$, where $x_{3,t} = (x'_{2,t}, w_t\ 1(t \geq T_B))'$ for w_t a scalar element of $x_{2,t}$ and $\beta_3 = (\beta'_2, \beta_{33})'$. Note that in this notation β^*_{33} takes the value of an element of $\Delta\beta^*_1$ if the test is for a break in the parameter associated with an element of the conditioning variables $x_{1,t}$ and takes the value of an element of $\Delta\beta^*_{22}$ if the test is for a break in the parameter associated with an element of the predictive variables $x_{22,t}$.

Proposition 1 (Chow test): The Chow test has noncentrality parameter

$$(1-\lambda_B)\left(\frac{[\lambda_B(\Delta\beta_2^{*\prime}Ex_{2,t}w_t - \beta_{33}^* Ew_t^2) + \beta_{33}^*(Ew_t^2 - (1-\lambda_B)(Ex_{2,t}w_t)'(Ex_{2,t}x_{2,t}')^{-1}(Ex_{2,t}w_t))]^2}{\sigma_u^2(Ew_t^2 - (1-\lambda_B)(Ex_{2,t}w_t)'(Ex_{2,t}x_{2,t}')^{-1}(Ex_{2,t}w_t))}\right).$$

There are a number of interesting aspects of this term that are relevant to us. First note that the noncentrality parameter is unaffected by breaks in the residual variance (embodied in the coefficient δ in our notation). We interpret this to mean not that there can be no effects on the power of the test but rather that those effects are of lower order importance than other aspects of the data generating process. The most important item of interest is that although the test can have power to detect breaks, whether or not that power is large depends in a complicated way upon a number of factors. Of particular note is that the test can be subject to both "spurious" and "hidden" breaks, because the noncentrality parameter depends not only on the break in the regression parameter being tested (i.e. $\beta_{33}^* = 0$) but also on all the other breaks that have occurred (those non-zero elements of $\Delta\beta_2^*$ not associated with w_t). A spurious break occurs when $\beta_{33}^* = 0$ but the noncentrality parameter is still positive. In this case the test will reject the null that there is no break in the particular regression coefficient of interest even if there is none. A hidden break occurs when $\beta_{33}^* \neq 0$ but other breaks interact to cause the noncentrality parameter to be zero (or small). In this case the test fails to reject the null that there is no break in the particular regression coefficient when there is one. It is for this reason that in Section 6 we note that in our empirical work we implement our (Andrews) break tests both with and without the conditioning on breaks in the other regression parameters.

Proposition 2 (Goldfeld-Quandt): The GQ test has noncentrality parameter $4(\delta\sigma_u^2)^2/V$.

The noncentrality parameter associated with the Goldfeld-Quandt test isn't nearly as complex as that for the Chow test. In particular whether or not it is non-zero is primarily determined by the magnitude of the variance break parameter δ. For our purposes this is important only in so far as we can express some confidence that our finding of breaks in the residual variance is not an artifact of unmodeled breaks in the regression parameters.

Proposition 3 (GC test): Let $A = Ex_{2,t}x_{2,t}' - Ex_{2,t}x_{1,t}'(Ex_{1,t}x_{1,t}')^{-1}Ex_{1,t}x_{2,t}'$. The GC test has noncentrality parameter $[\beta_{22}^* + (1-\lambda_B)\Delta\beta_{22}^*]'A[\beta_{22}^* + (1-\lambda_B)\Delta\beta_{22}^*]/\sigma_u^2$.

Proposition 3 gives us our first glimpse into the asymptotic power of tests of predictive ability. As was the case for the Chow test, the noncentrality parameter provides a number of interesting items of note. In particular we find that when using the standard in-sample test of predictive ability, breaks in the parameters associated with the conditioning variables and breaks in the residual variance are asymptotically irrelevant. Again we interpret this not as meaning that such breaks can have no effect but rather that their effect is of lower order importance than other aspects of the data generating process. Even so, this is important for our analysis of the predictive content of the output gap since our empirical work indicates the potential for breaks in the parameters associated with the lagged changes in inflation and the residual variance. The absence of a dependence of the causality test's noncentrality parameter on $\Delta\beta_1^*$ or δ lends support to our argument that the out-of-sample breakdown in the predictive content of the output gap for inflation is not driven by breaks in the parameters associated with the conditioning variables or the residual variance.

It is also important to note that structural change in the parameters associated with the output gap would likely not change whether the in-sample F-test indicated predictive ability. So long as $\beta_{22}^* + (1-\lambda_B)\Delta\beta_{22}^* \neq 0$ the F-test will reject with probability one in large samples. This implies that unmodeled breaks in the parameters associated with the output gap are unlikely to have an effect on whether or not the F-test rejects. Again this is important for our empirical analysis since we find that although graphically there appears to be substantial variation in the parameter estimates associated with the output gap, this variation is insufficient to lead standard break tests to reject the null of no structural change. We interpret this result as indicating that the output gap does contain predictive content for inflation though perhaps not on a consistent basis. This possibility is made clearer using the two out-of-sample tests of predictive ability.

Proposition 4 (MSE-F): Let $A = Ex_{2,t}x_{2,t}' - Ex_{2,t}x_{1,t}'(Ex_{1,t}x_{1,t}')^{-1}Ex_{1,t}x_{2,t}'$. (a) If $T_B < R$, MSE-F has noncentrality parameter $(1-\lambda)\{\beta_{22}^{*\prime}A\beta_{22}^* + 2\beta_{22}^{*\prime}A\Delta\beta_{22}^* + (1-\lambda_B^2/\lambda)\Delta\beta_{22}^{*\prime}A\Delta\beta_{22}^*\}/\sigma_u^2$, (b) If $T_B \geq R$, MSE-F has noncentrality parameter $\{(1-\lambda)\beta_{22}^{*\prime}A\beta_{22}^* + 2(1-\lambda_B)\beta_{22}^{*\prime}A\Delta\beta_{22}^* + (1-\lambda_B)^2\Delta\beta_{22}^{*\prime}A\Delta\beta_{22}^*\}/\sigma_u^2$.

Proposition 5 (ENC-F): Let $A = Ex_{2,t}x_{2,t}' - Ex_{2,t}x_{1,t}'(Ex_{1,t}x_{1,t}')^{-1}Ex_{1,t}x_{2,t}'$. (a) If

$T_B < R$, ENC-F has noncentrality parameter $\{(1-\lambda)\beta_{22}^{*\prime}A\beta_{22}^{*}+(2(1-\lambda)+\lambda_B \ln \lambda_B)\beta_{22}^{*\prime}A\Delta\beta_{22}^{*}+(1-\lambda+\lambda_B \ln \lambda_B)\Delta\beta_{22}^{*\prime}A\Delta\beta_{22}^{*}\}/\sigma_u^2$, (b) If $T_B \geq R$, MSE-F has noncentrality parameter $\{(1-\lambda)\beta_{22}^{*\prime}A\beta_{22}^{*}+(2(1-\lambda_B)+\lambda_B \ln \lambda_B)\beta_{22}^{*\prime}A\Delta\beta_{22}^{*}+(1-\lambda_B+\lambda_B \ln \lambda_B)\Delta\beta_{22}^{*\prime}A\Delta\beta_{22}^{*}\}/\sigma_u^2$.

Propositions 4 and 5 provide the noncentrality parameters associated with the two out-of-sample tests of predictive ability. As was the case for the in–sample GC test, we find that breaks in the parameters associated with the conditioning variables and breaks in the residual variance are asymptotically irrelevant. Again we interpret this not as meaning that such breaks can have no effect but rather that their effect is of lower order importance than other aspects of the data generating process. Even so, the absence of a dependence of the noncentrality parameter on $\Delta\beta_1^*$ or δ lends support to our argument that the out-of-sample breakdown in the predictive content of the output gap for inflation is not driven by breaks in the parameters associated with the conditioning variables or the residual variance.

What makes these noncentrality parameters most interesting, relative to that for the GC test, is that they can take both positive and negative values. Recall that the upper tail of the null distribution of these tests is used as the critical region. This is a natural choice since if $\beta_{22}^* \neq 0$ and $\Delta\beta_{22}^* = 0$ (so the DGP is stable and includes the output gap as a regressor), both noncentrality parameters are strictly positive. This argument can fail when there are structural breaks. For a given value of β_{22}^*, structural breaks (i.e. $\Delta\beta_{22}^* \neq 0$) can occur that cause either of the noncentrality parameters to be negative and hence we would "fail to reject" the null that the output gap exhibits predictive ability for inflation. This implies that unmodeled breaks in the parameters associated with the output gap can very well be an explanation for the finding of in-sample but not out-of-sample predictive ability of the output gap for inflation.

9 Conclusions

In this paper we have sought to explain how the seemingly good in–sample fits of Phillips curve models fail to translate into consistently good out-of-sample forecast performance. We focus on models relating inflation to the output gap because this particular formulation seems to have garnered the greatest interest in recent years, in part because of the development of dynamic, stochastic, general equilibrium models that yield a Phillips curve formulation relating inflation to expected future inflation and the output gap.

After presenting our own sample forecast evidence for a range of inflation and output gap measures, we use Monte Carlo simulations to examine the extent to which the in–sample and out–of–sample evidence can be explained by the potentially lower power of out–of–sample metrics or instabilities in the model. In particular, we compare our sample estimates to results of Monte Carlo simulations of empirically–based data–generating processes that (i) impose stability or (ii) allow various types of discrete breaks in the model parameters. The imposed shifts in the DGPs are based on the results of extensive break test analysis applied to the empirical models in question. The break tests provide strong evidence of shifts in the residual variances of the inflation equations, modest evidence of shifts in the AR parameters of the inflation equations, and only very limited evidence of shifts in the output gap coefficients.

Our simulation results show that, while power differences can account for some portion of the gap between the in–sample and out–of–sample success of the Phillips curve, much of the gulf can be accounted for by instabilities in the coefficients on the output gap. While formal break tests yield little evidence of such breaks, the simulations show the power of the tests is typically only modest. The types of breaks for which the formal test evidence is considerably stronger — breaks in residual variances and the AR parameters of the inflation equations — generally fail to generate a breakdown in the predictive content of the output gap.

Building on the work of Clark and McCracken (2002b), we conclude the paper by showing analytically that breaks in the output gap coefficients, but not breaks in residual variances or AR coefficients, lead to a breakdown in the power of tests of equal forecast accuracy and forecast encompassing. We do so using a local alternatives framework, like that of Hansen (2000).

In light of our findings, an important outstanding question — beyond the scope of this paper — is, what forecast methods work well in the face of instability? One approach, considered in such studies as Stock and Watson (1996), Canova (2001), and Marcellino (2002), is to allow time-varying parameters. But consistent with the mixed existing evidence on the efficacy of TVP, we have found that allowing TVP fails to improve the forecast performance of the Phillips curve models considered in this paper. Alternatively, based on a wide battery of results, Stock and Watson (2001, 2003) have suggested that forecast combination — a particular form of shrinkage — may be a way of overcoming instabilities.

Other shrinkage approaches could also be useful. Finally, Pesaran and Timmermann (2002) have proposed a two-step method of working backward in time to identify the most recent break and then using just the post-break data to estimate a model and forecast.

References

Andrews, Donald W.K., 1993, "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica* 61, 821-56.

Atkeson, Andrew, and Lee E. Ohanian, 2001, "Are Phillips Curves Useful for Forecasting Inflation?," *Federal Reserve Bank of Minneapolis Quarterly Review*, v.25, pp.2-11.

Bai, Jushan, and Pierre Perron, 1998, "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica* 66, 47-78.

Bai, Jushan, and Pierre Perron, 2002, "Computation and Analysis of Multiple Structural-Change Models," *Journal of Applied Econometrics*, forthcoming.

Blinder, Alan S., 1997, "Is There a Core of Practical Macroeconomics That We Should All Believe?," *American Economic Review* 87 (May), pp.240-43.

Brayton, Flint, John M. Roberts, and John C. Williams, 1999, "What's Happened to the Phillips Curve?," manuscript, Board of Governors of the Federal Reserve System.

Camba–Mendez, Gonzalo, and Diego Rodriguez–Palenquela, 2001, "Assessment Criteria for Output Gap Estimates," Working Paper No. 54, European Central Bank.

Canova, Fabio, 2001, "G-7 Inflation Forecasts," manuscript, Universitat Pompeu Fabra.

Claus, Iris, 2000, "Is the Output Gap a Useful Indicator of Inflation?," DP2000/05, Reserve Bank of New Zealand.

Cecchetti, Stephen G., 1995, "Inflation Indicators and Inflation Policy," *NBER Macroeconomics Annual*, pp. 189-219.

Clarida, Richard H., Lucio Sarno, Mark P. Taylor, and Giorgio Valente, 2001, "The Out–of–Sample Success of Term Structure Models as Exchange Rate Predictors: A Step Beyond," NBER Working Paper #8601.

Clark, Todd E., and Michael W. McCracken, 2001(a), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105 (November), pp. 85-110.

Clark, Todd E., and Michael W. McCracken, 2002(a), "Evaluating Long–Horizon Forecasts," manuscript, Federal Reserve Bank of Kansas City.

Clark, Todd E., and Michael W. McCracken, 2002(b), "Forecast-Based Model Selection in the Presence of Structural Breaks," manuscript, Federal Reserve Bank of Kansas City.

Clark, Todd E., and Michael W. McCracken, 2003, "Technical Appendix to 'The Predictive Content of the Output Gap for Inflation: Resolving In–Sample and Out–of–Sample

Evidence',” manuscript, Federal Reserve Bank of Kansas City.

Clements, Michael P., 2002, “Why Forecast Performance Does Not Help Us Choose a Model,” manuscript, University of Warwick.

Diebold, Francis X., and Celia Chen, 1996, “Testing Structural Stability with Endogenous Breakpoint: A Size Comparison of Analytic and Bootstrap Procedures,” *Journal of Econometrics* 70 (January), pp. 221-41.

Diebold, Francis X., and Roberto S. Mariano, 1995, “Comparing Predictive Accuracy,” *Journal of Business and Economic Statistics* 13 (July), pp. 253-63.

Fisher, Paul G., Lavan Mahadeva, and John D. Whitley, 1997, “The Output Gap and Inflation – Experience at the Bank of England,” *BIS Conference Papers* 4, pp. 68-90.

Fisher, Jonas D.M., C.T. Liu, and R. Zhou, 2002, “When Can We Forecast Inflation?,” *Economic Perspectives*, Federal Reserve Bank of Chicago (First Quarter), pp. 30-42.

Gali, Jordi, and Mark Gertler, 1999, “Inflation Dynamics: A Structural Econometric Analysis,” *Journal of Monetary Economics* 44, pp. 195-222.

Gali, Jordi, Mark Gertler, and J. David Lopez-Salido, 2001, “European Inflation Dynamics,” *European Economic Review* 45, pp. 1237-70.

Gerlach, Stefan, and Frank Smets, 1999, “Output Gaps and Monetary Policy in the EMU Area,” *European Economic Review* 43, pp.801-12.

Gordon, Robert J., 1990, “What Is New–Keynesian Economics?” *Journal of Economic Literature* 28 (Sept.), pp. 1115-1171.

Hansen, Bruce E., 1997, “Approximate Asymptotic P Values for Structural-Change Models,” *Journal of Business and Economic Statistics* 15, 60-67.

Hansen, Bruce E., 2000, “Testing for Structural Change in Conditional Models,” *Journal of Econometrics* 97, 93-116.

Hansen, Bruce E., 2001, “The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity,” *Journal of Economic Perspectives* 15 (Fall), p. 117-28.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold, 1998, “Tests for Forecast Encompassing,” *Journal of Business and Economic Statistics* 16 (April), pp. 254-59.

Hodrick, Robert J., and Edward C. Prescott, 1997, “Postwar U.S. Business Cycles: an Empirical Investigation,” *Journal of Money, Credit, and Banking* 29, 1-16.

Inoue, Atsushi, and Lutz Kilian, 2002, “In-Sample or Out-of-Sample Tests of Predictability? Which One Should We Use?,” manuscript, North Carolina State University.

Kilian, Lutz, 1999, "Exchange Rates and Monetary Fundamentals: What Do We Learn From Long–Horizon Regressions?," *Journal of Applied Econometrics* 14 (Sept.-Oct.), pp. 491-510.

Kilian, Lutz, and Mark P. Taylor, 2001, "Why Is It So Difficult to Beat the Random Walk Forecast of Exchange Rates?" manuscript, University of Michigan.

King, Robert G., 2000, "The New IS-LM Model: Language, Logic, and Limits," *Economic Quarterly*, Federal Reserve Bank of Richmond (Summer), pp.45-103.

King, Robert G., and Mark W. Watson, 1994, "The Postwar U.S. Phillips Curve: A Revisionist Econometric History," *Carnegie–Rochester Conference Series on Public Policy* 41 (December), pp.157-219.

Kuttner, Kenneth N., 1994, "Estimating Potential Output as a Latent Variable," *Journal of Business and Economic Statistics* 12, pp.361-68.

Lown, Cara S., and Robert W. Rich, 1997, "Is there an inflation puzzle?," *Economic Policy Review*, Federal Reserve Bank of New York (Dec.), pp. 51-77.

Marcellino, Massimiliano, 2002, "Instability and Non-Linearity in the EMU," IGIER Working Paper no. 211.

Mark, Nelson C., 1995, "Exchange Rates and Fundamentals: Evidence on Long–Horizon Predictability," *American Economic Review* 85 (March), pp. 201-18.

Mark, Nelson C., and Donggyu Sul, 2002, "Asymptotic Power Advantages of Long–Horizon Regression Tests," manuscript, Ohio State University.

McCallum, Bennett T., and Edward Nelson, 2001, "An Optimizing IS-LM Specification for Monetary Policy and Business Cycle Analysis," *Journal of Money, Credit, and Banking* 31, pp.296-316

McCracken, Michael W., 2000, "Asymptotics for Out–of–Sample Tests of Causality," manuscript, University of Missouri.

Meese, Richard A. and Kenneth Rogoff, 1983, "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?", *Journal of International Economics* 14, pp. 3-24.

Meese, Richard, and Kenneth Rogoff, 1988, "Was It Real? The Exchange Rate–Interest Differential Relation Over The Modern Floating–Rate Period," *Journal of Finance* 43 (September), pp. 933-948.

Newey, Whitney K., and Kenneth D. West, 1987, "A Simple, Positive Semi-definite, Het-

eroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55 (May), pp. 703-08.

Nyblom, Jukka, 1989, "Testing for the Constancy of Parameters Over Time," *Journal of the American Statistical Association* 84, 223-230.

Orphanides, Athanasios, 2000, "The Quest for Prosperity Without Inflation," Working Paper No. 15, European Central Bank.

Orphanides, Athanasios, and Simon van Norden, 2001, "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time," manuscript, CIRANO.

Peersman, Gert, and Frank Smets, 1999, "Uncertainty and the Taylor Rule in a Simple Model of the Euro-Area Economy," manuscript.

Pesaran, M. Hashem, and Allan Timmermann, 2002, "Market Timing and Return Prediction under Model Instability," *Journal of Empirical Finance*, forthcoming.

Reifschnieder, David, Robert Tetlow, and John Williams, 1999, "Aggregate Disturbances, Monetary Policy, and the Macroeconomy: The FRB/US Perspective," *Federal Reserve Bulletin* 85 (Jan.), pp. 1-19.

Roberts, John M., 1995, "New Keynesian Economics and the Phillips Curve," *Journal of Money, Credit, and Banking* 27 (Nov.), pp. 975-84.

Rogerson, Richard, 1997, "Theory Ahead of Language in the Economics of Unemployment," *Journal of Economic Perspectives* 11 (Winter), pp. 73-92.

Rotemberg, Julio J., and Michael Woodford, 1997, "An Optimization–Based Econometric Framework for the Evaluation of Monetary Policy," *NBER Macroeconomics Annual*, pp. 297-346.

Rudebusch, Glenn D., 2002, "Assessing the Lucas Critique in Monetary Policy Models," manuscript, Federal Reserve Bank of San Francisco.

Rudebusch, Glenn D., and Lars E.O. Svensson, 1999, "Policy Rules for Inflation Targeting," in *Monetary Policy Rules*, John B. Taylor, ed., Chicago: University of Chicago Press, pp. 203-46.

Sbordone, Argia M., 2002, "Prices and Unit Labor Costs: A New Test of Price Stickiness," *Journal of Monetary Economics* 49(2), 265-92.

Staiger, Douglas, James H. Stock and Mark W. Watson, 1997, "The NAIRU, Unemployment and Monetary Policy," *Journal of Economic Perspectives* 11 (Winter), pp. 33-49.

Staiger, Douglas, James H. Stock and Mark W. Watson, 2001, "Prices, Wages, and the U.S.

NAIRU in the 1990s," NBER Working Paper #8320.

St-Amant, Pierre, and Simon van Norden, 1997, Measurement of the Output Gap: A Discussion of Recent Research at the Bank of Canada," *BIS Conference Papers* 4, pp. 1-38.

Stine, Robert A., 1987, "Estimating Properties of Autoregressive Forecasts," *Journal of the American Statistical Association* 82 (December), pp. 1072-78.

Stock, James H., and Mark W. Watson, 1996, "Evidence on Structural Stability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics* 14, pp. 11-30.

Stock, James H., and Mark W. Watson, 1999(a), "Business Cycle Fluctuations in U.S. Macroeconomic Time Series," in *Handbook of Macroeconomics*, Volume 1, J. Taylor and M. Woodford., eds., (North Holland).

Stock, James H., and Mark W. Watson, 1999 (b), "Forecasting Inflation," *Journal of Monetary Economics* 44 (October), pp. 293-335.

Stock, James H., and Mark W. Watson, 2001, "Forecasting Output and Inflation: The Role of Asset Prices," NBER Working Paper #8180.

Stock, James H., and Mark W. Watson, 2002, "Has the Business Cycle Changed and Why?," *NBER Macroeconomics Annual*.

Stock, James H., and Mark W. Watson, 2003, "Combination Forecasts of Output Growth in a Seven–Country Data Set," manuscript, Princeton University.

West, Kenneth D., 1996, "Asymptotic Inference About Predictive Ability," *Econometrica* 64 (September), pp. 1067-84.

West, Kenneth D., 2001, "Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters," *Journal of Business and Economic Statistics* 19 (January), pp. 29-33.

Table 1: Full–Sample Model Estimates

Dependent variable: $\pi_{t+\tau}^{(\tau)} - \pi_t$

Sample: 1958:3+$\tau - 1$ *to 2001:4*

	GDP price index			
	CBO Gap		**HP Gap**	
	coef. (s.e.)		*coef. (s.e.)*	
	$\tau = 1$	$\tau = 4$	$\tau = 1$	$\tau = 4$
constant	.023 (.074)	.076 (.111)	-.012 (.074)	.008 (.116)
$\Delta\pi_t$	-.332 (.093)	-.439 (.059)	-.338 (.090)	-.432 (.067)
$\Delta\pi_{t-1}$	-.279 (.074)	-.294 (.061)	-.288 (.074)	-.290 (.080)
gap_t	.128 (.030)	.223 (.056)	.223 (.053)	.341 (.101)
$\bar{R}^2$	.154	.318	.165	.284
	Core CPI			
	CBO Gap		**HP Gap**	
	coef. (s.e.)		*coef. (s.e.)*	
	$\tau = 1$	$\tau = 4$	$\tau = 1$	$\tau = 4$
constant	.051 (.094)	.093 (.155)	.009 (.098)	.014 (.153)
$\Delta\pi_t$	-.392 (.135)	-.482 (.108)	-.423 (.134)	-.515 (.098)
$\Delta\pi_{t-1}$	-.292 (.119)	-.311 (.080)	-.327 (.120)	-.350 (.084)
gap_t	.158 (.052)	.261 (.085)	.304 (.083)	.450 (.147)
$\bar{R}^2$	.167	.272	.189	.280

Notes:
1. As indicated in equation (1), $\pi_t^{(\tau)} \equiv (400/\tau) * \ln(p_t/p_{t-\tau})$.
2. The reported standard errors are robust with respect to heteroskedasticity.

Table 2: Sample Forecasting Results for the GDP Price Index

	CBO Gap						**HP Gap**					
	RMSEs, τ = 1						*RMSEs, τ = 1*					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
RMSE ratio	0.973	0.976	0.968	0.973	0.947	0.995	1.011	1.072	0.923	1.092	0.969	0.937
RMSE,AR	0.884	0.924	0.839	0.989	0.840	0.822	0.884	0.924	0.839	0.989	0.840	0.822
	Bootstrapped p-values of tests, τ = 1						*Bootstrapped p-values of tests, τ = 1*					
MSE-F	0.002	0.021	0.004	0.031	0.004	0.170	0.834	0.998	0.000	0.994	0.022	0.000
MSE-t	0.039	0.136	0.088	0.172	0.051	0.275	0.274	0.764	0.010	0.758	0.148	0.040
ENC-F	0.000	0.001	0.003	0.000	0.004	0.021	0.000	0.018	0.000	0.025	0.007	0.000
ENC-t	0.001	0.026	0.031	0.026	0.032	0.138	0.026	0.152	0.006	0.165	0.061	0.028
	RMSEs, τ = 4						*RMSEs, τ = 4*					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
RMSE ratio	0.905	0.894	0.946	0.897	0.841	1.070	1.082	1.115	0.946	1.108	1.079	0.922
RMSE,AR	0.890	1.115	0.575	1.313	0.757	0.484	0.890	1.115	0.575	1.313	0.757	0.484
	Bootstrapped p-values of tests, τ = 4						*Bootstrapped p-values of tests, τ = 4*					
MSE-F	0.001	0.010	0.032	0.023	0.005	0.904	0.984	0.970	0.027	0.927	0.932	0.017
MSE-t	0.069	0.137	0.190	0.196	0.053	0.474	0.631	0.801	0.155	0.700	0.538	0.095
ENC-F	0.000	0.001	0.000	0.003	0.005	0.003	0.003	0.022	0.002	0.030	0.083	0.002
ENC-t	0.030	0.082	0.017	0.098	0.040	0.052	0.050	0.112	0.017	0.115	0.237	0.009

Notes:
1. The *RMSE ratio* is the ratio of the RMSE of the model including the output gap (equation (1)) to the RMSE of the AR model (equation (2)), which is reported in the row labeled *RMSE,AR*. τ refers to the forecast horizon.
2. The test statistics MSE-F, MSE-t, ENC-F, and ENC-t are defined in section 3.2. The bootstrap procedure used to construct p–values is detailed in section 3.3.

Table 3: Sample Forecasting Results for the Core CPI

	CBO Gap						**HP Gap**					
	RMSEs, τ = 1						*RMSEs, τ = 1*					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
RMSE ratio	0.967	0.966	0.976	0.964	0.960	1.105	0.958	0.960	0.926	0.951	1.025	1.009
RMSE,AR	1.527	2.055	0.534	2.567	0.740	0.364	1.527	2.055	0.534	2.567	0.740	0.364
	Bootstrapped p-values of tests, τ = 1						*Bootstrapped p-values of tests, τ = 1*					
MSE-F	0.001	0.008	0.019	0.015	0.011	0.999	0.000	0.005	0.000	0.007	0.908	0.746
MSE-t	0.048	0.082	0.207	0.098	0.127	0.641	0.041	0.084	0.082	0.084	0.411	0.335
ENC-F	0.002	0.005	0.000	0.012	0.003	0.001	0.000	0.002	0.000	0.007	0.010	0.000
ENC-t	0.040	0.068	0.023	0.062	0.071	0.100	0.027	0.051	0.008	0.054	0.137	0.022
	RMSEs, τ = 4						*RMSEs, τ = 4*					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
RMSE ratio	0.953	0.954	0.930	0.955	0.787	1.338	0.998	1.006	0.870	0.978	1.230	0.925
RMSE,AR	1.416	1.925	0.517	2.465	0.673	0.325	1.416	1.925	0.517	2.465	0.673	0.325
	Bootstrapped p-values of tests, τ = 4						*Bootstrapped p-values of tests, τ = 4*					
MSE-F	0.021	0.059	0.021	0.096	0.001	0.998	0.168	0.384	0.002	0.165	0.990	0.024
MSE-t	0.129	0.192	0.222	0.236	0.072	0.807	0.185	0.293	0.076	0.258	0.682	0.200
ENC-F	0.003	0.013	0.000	0.026	0.001	0.004	0.003	0.025	0.000	0.033	0.134	0.000
ENC-t	0.086	0.136	0.026	0.159	0.067	0.058	0.057	0.117	0.012	0.062	0.368	0.002

Notes:
1. See the notes to Table 2.

Table 4: Simulated Power Results from Stable DGPs for the GDP Price Index

	CBO Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.908	.756	.792	.667	.695	.737	.877	.714	.752	.635	.656	.722
MSE-t	.870	.635	.598	.498	.481	.508	.811	.573	.556	.447	.420	.458
ENC-F	.966	.883	.921	.788	.844	.897	.949	.835	.898	.739	.803	.875
ENC-t	.948	.812	.815	.672	.673	.736	.923	.714	.750	.560	.563	.649
RMSE ratio results:												
median	.962	.967	.964	.968	.967	.965	.902	.913	.900	.913	.915	.902
prob.<1	.927	.834	.856	.761	.767	.814	.902	.801	.815	.763	.763	.796
5%ile	.903	.890	.879	.874	.860	.870	.754	.718	.711	.681	.662	.693
95%ile	1.005	1.025	1.024	1.044	1.051	1.030	1.025	1.098	1.080	1.181	1.158	1.114
full-sample t-test	.990						.984					
	HP Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.951	.831	.813	.730	.753	.763	.900	.768	.757	.664	.722	.714
MSE-t	.910	.705	.627	.563	.519	.510	.831	.598	.549	.447	.458	.430
ENC-F	.988	.948	.954	.874	.908	.933	.964	.873	.910	.782	.852	.866
ENC-t	.983	.880	.842	.744	.751	.764	.937	.739	.756	.573	.639	.633
RMSE ratio results:												
median	.955	.957	.954	.955	.955	.957	.912	.918	.911	.916	.908	.914
prob.<1	.959	.871	.861	.811	.825	.833	.931	.852	.836	.787	.813	.800
5%ile	.898	.883	.874	.853	.860	.865	.793	.761	.753	.713	.738	.732
95%ile	.996	1.017	1.021	1.037	1.035	1.032	1.006	1.064	1.059	1.115	1.101	1.099
full-sample t-test	.998						.990					

Notes:
1. The data generating process is based on empirical estimates of equation (1) for the forecast horizon $\tau = 1$ and a simple AR model for the output gap. The inflation equations are given in Table 1. All of the DGPs suppose stability in the coefficients and residual variances. In each of these "true" models, the output gap has predictive power for inflation, in the sense of having a significant regression coefficient.

2. For each artificial data set, forecasts of $\pi^{(\tau)}_{t+\tau} - \pi_t$ are formed recursively using estimates of equations (1) and (2). These forecasts are then used to form the indicated test statistics, defined in Section 3.2, as well as the ratio of the RMSE of the model including the output gap to the RMSE of the AR model. In each Monte Carlo replication, the simulated test statistics are compared against bootstrapped null critical values (with a significance level of 10%), computed using the procedure described in section 3.3.

3. For the test statistics of interest, the table reports the percentage of Monte Carlo simulations in which the null of equal MSE or forecast encompassing is rejected. For the RMSE ratio, the table reports several summary statistics for the distribution of the ratio in the Monte Carlo trials: the median; the percent of the Monte Carlo ratios that are less than 1 (labeled *prob.<1*); and a 90% interval (given by the *5%ile* and *95%ile* figures). *Full-sample t-test* refers to the usual t–test of the null that the output gap coefficient is less than or equal to zero.

4. The number of Monte Carlo simulations is 1000; the number of bootstrap draws is 2000.

Table 5: Simulated Power Results from Stable DGPs for the Core CPI

	CBO Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.858	.739	.712	.625	.635	.637	.827	.674	.695	.548	.620	.629
MSE-t	.810	.616	.535	.455	.453	.442	.753	.528	.476	.387	.410	.388
ENC-F	.931	.842	.854	.739	.780	.819	.916	.788	.837	.664	.756	.783
ENC-t	.913	.767	.730	.619	.612	.631	.878	.663	.662	.498	.523	.540
RMSE ratio results:												
median	.970	.970	.972	.975	.972	.974	.919	.930	.927	.938	.929	.932
prob.<1	.890	.804	.783	.733	.746	.751	.866	.787	.786	.725	.729	.757
5%ile	.912	.897	.889	.880	.876	.883	.788	.756	.750	.722	.709	.724
95%ile	1.010	1.023	1.030	1.043	1.044	1.037	1.033	1.079	1.096	1.148	1.136	1.106
full-sample t-test	.978						.966					
	HP Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.937	.823	.836	.723	.759	.776	.889	.741	.765	.633	.697	.711
MSE-t	.887	.702	.636	.541	.548	.514	.823	.575	.523	.432	.437	.437
ENC-F	.987	.941	.945	.860	.905	.928	.959	.856	.913	.760	.837	.870
ENC-t	.975	.877	.872	.739	.757	.785	.925	.731	.746	.542	.584	.624
RMSE ratio results:												
median	.953	.957	.954	.958	.954	.956	.917	.921	.916	.926	.912	.915
prob.<1	.954	.869	.874	.808	.827	.828	.923	.827	.838	.756	.798	.793
5%ile	.900	.873	.877	.852	.863	.865	.803	.758	.762	.720	.720	.745
95%ile	.999	1.018	1.024	1.046	1.034	1.031	1.012	1.066	1.059	1.125	1.106	1.097
full-sample t-test	.997						.993					

Notes:
1. See the notes to Table 4.

Table 6: Break Test Results, $\tau = 1$

Asymptotic p–values (break dates)

	GDP price index		Core CPI		Core PCE index	
	CBO Gap	*HP Gap*	*CBO Gap*	*HP Gap*	*CBO Gap*	*HP Gap*
constant	.74	.20	.41	.39	.94	.36
$\Delta\pi$ coefs.	.46	.57	.02 (80:2)	.03 (80:1)	.06 (83:2)	.07 (83:2)
gap_{t-1}	.59	.27	.05 (83:1)	.01 (83:1)	.73	.33
σ^2: homosk.	.24	.18	.01 (84:1)	.02 (83:4)	.26	.26
σ^2: heterosk.	.00 (66:4)	.00 (66:4)	.00 (95:1)	.00 (94:1)	.00 (66:4)	.00 (66:4)

Notes:
1. The table reports asymptotic p–values for Andrews' (1993) SupWald test applied to the coefficients and residual variance and estimates of equation (1) with $\tau = 1$ (regression estimates supposing stability are given in Table 1). Tests are applied separately to the intercept, the set of coefficients on lagged inflation, the coefficient on the output gap, and the residual variance. In testing the stability of a given parameter or set of coefficients, the other coefficients in the model are assumed stable. The p–values are computed using Hansen's (1997) approximation.
2. In those cases in which the SupWald test indicates a significant (at 10% or less) break, the table reports the break date estimate (simply the date of the maximum of the sequence of Wald tests).
3. The Wald tests are computed with heteroskedasticity–robust variance estimates. In the case of the tests for the residual variance, the table reports results both allowing for heteroskedasticity and imposing homoskedasticity.

Table 7: Tests of Multiple Breaks in Residual Variances, $\tau = 1$

	GDP price index		*Core CPI*	
	CBO Gap	*HP Gap*	*CBO Gap*	*HP Gap*
SupF(1)	5.323	5.317	11.451 **	9.741 **
SupF(2)	15.517 ***	12.784 ***	20.945 ***	17.051 ***
SupF(3)	10.652 ***	9.009 ***	14.006 ***	11.407 ***
UDmax	15.517 ***	12.784 ***	20.945 ***	17.051 ***
WDmax(1%)	20.374 ***	16.786 ***	27.501 ***	22.388 ***
SupF(2|1)	9.332 *	8.353	15.363 ***	17.309 ***
SupF(3|2)	1.614	2.815	14.962 ***	20.373 ***

Notes:
1. The table reports the results of Bai and Perron (1998, 2002) tests for multiple breaks applied to the residual variances of the inflation equation (1) with $\tau = 1$. The squared regression residuals on which the tests are based are taken from models that impose the breaks in the regression coefficients described in section 7.2.
2. The *SupF(i)* statistics are tests of the null of no breaks against the alternative of i breaks. The UDmax and WDmax statistics are general tests for the presence of some number of breaks. The *SupF(i+1\|i)* statistics are tests of the null of i breaks against the alternative of $i+1$ breaks. ***, **, and * denote statistics that are statistically significant at the 1%, 5%, and 10% levels, respectively.
3. The reported test statistics use the HAC correction built into the Bai–Perron computer code.

Table 8: Estimates of Models with Breaks in AR Coefficients

Dependent variable: $\Delta\pi_t$

Sample: 1958:3 to 2001:4

	GDP price index		**Core CPI**	
	coef. (s.e.)		*coef. (s.e.)*	
	CBO Gap	*HP Gap*	*CBO Gap*	*HP Gap*
constant	.017 (.073)	-.016 (.073)	-.015 (.088)	-.052 (.095)
$\Delta\pi_{t-1}$, $t \leq BP_\pi$	-.259 (.118)	-.268 (.115)	-.107 (.145)	-.151 (.143)
$\Delta\pi_{t-2}$, $t \leq BP_\pi$	-.259 (.090)	-.269 (.093)	-.038 (.140)	-.085 (.135)
$\Delta\pi_{t-1}$, $t > BP_\pi$	-.548 (.098)	-.543 (.100)	-.661 (.142)	-.659 (.149)
$\Delta\pi_{t-2}$, $t > BP_\pi$	-.371 (.110)	-.370 (.110)	-.538 (.154)	-.542 (.154)
gap_{t-1}	.126 (.030)	.218 (.052)	.158 (.052)	.275 (.077)
$\bar{R}^2$	.159	.169	.261	.266

Notes:
1. The table reports estimates of equation (1) allowing a single break in the set of coefficients on lagged inflation. The break date BP_π is specified as 1983:1 for the GDP price index and 1980:2 for the core CPI.
2. The reported standard errors are robust with respect to heteroskedasticity.

Table 9: Simulated Power Results for the GDP Price Index with Breaks in AR Coefs.

	CBO Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.930	.856	.818	.789	.753	.771	.866	.789	.720	.728	.661	.689
MSE-t	.895	.761	.650	.605	.552	.551	.808	.642	.534	.501	.443	.469
ENC-F	.980	.953	.920	.911	.863	.881	.961	.915	.863	.857	.812	.827
ENC-t	.975	.905	.809	.808	.694	.723	.935	.817	.737	.679	.617	.645
RMSE ratio results:												
median	.958	.953	.968	.945	.967	.969	.894	.876	.919	.862	.919	.921
prob.<1	.955	.898	.889	.855	.855	.852	.897	.842	.801	.815	.770	.795
5%ile	.897	.861	.906	.828	.890	.903	.727	.651	.728	.605	.684	.720
95%ile	.999	1.015	1.011	1.035	1.022	1.020	1.041	1.076	1.083	1.143	1.127	1.104
full-sample t-test	.990						.971					
	HP Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.957	.883	.837	.810	.809	.762	.879	.785	.704	.709	.658	.659
MSE-t	.930	.790	.680	.616	.596	.575	.803	.602	.511	.457	.448	.437
ENC-F	.983	.966	.932	.947	.902	.900	.961	.919	.864	.872	.820	.806
ENC-t	.977	.917	.826	.842	.746	.716	.934	.800	.698	.630	.611	.604
RMSE ratio results:												
median	.953	.945	.963	.937	.961	.963	.913	.898	.938	.885	.926	.938
prob.<1	.975	.917	.890	.862	.898	.838	.907	.853	.789	.815	.772	.761
5%ile	.900	.865	.912	.822	.897	.905	.784	.720	.796	.656	.757	.776
95%ile	.994	1.009	1.014	1.034	1.015	1.023	1.020	1.065	1.085	1.123	1.113	1.101
full-sample t-test	.994						.977					

Notes:
1. The data generating process is based on empirical estimates of equation (1) for the forecast horizon $\tau = 1$ and a simple AR model for the output gap, allowing breaks in the coefficients on lagged inflation in the inflation equation and in the residual variance–covariance matrix. The break in the inflation coefficients occurs in 1983:1. The residual variance of the inflation equation has breaks in 1970:2 and 1976:2; the variance of the gap equation has a break in 1984:1. The inflation equations are given in Table 8. In each of these "true" models, the output gap has predictive power for inflation, in the sense of having a significant regression coefficient.

2. For each artificial data set, forecasts of $\pi_{t+\tau}^{(\tau)} - \pi_t$ are formed recursively using estimates of equations (1) and (2). These forecasts are then used to form the indicated test statistics, defined in Section 3.2, as well as the ratio of the RMSE of the model including the output gap to the RMSE of the AR model. In each Monte Carlo replication, the simulated test statistics are compared against bootstrapped null critical values (with a significance level of 10%), computed using the procedure described in section 3.3.

3. For the test statistics of interest, the table reports the percentage of Monte Carlo simulations in which the null of equal MSE or forecast encompassing is rejected. For the RMSE ratio, the table reports several summary statistics for the distribution of the ratio in the Monte Carlo trials: the median; the percent of the Monte Carlo ratios that are less than 1 (labeled *prob.<1*); and a 90% interval (given by the *5%ile* and *95%ile* figures). *Full-sample t-test* refers to the usual t–test of the null that the output gap coefficient is less than or equal to zero.

4. The number of Monte Carlo simulations is 1000; the number of bootstrap draws is 2000.

Table 10: Simulated Power Results for the Core CPI with Breaks in AR Coefs.

	CBO Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.828	.716	.852	.619	.765	.821	.735	.631	.716	.518	.661	.702
MSE-t	.766	.607	.672	.488	.525	.596	.654	.464	.553	.362	.435	.498
ENC-F	.889	.773	.976	.672	.916	.975	.868	.744	.951	.626	.895	.934
ENC-t	.806	.660	.921	.573	.760	.875	.784	.643	.872	.533	.679	.787
RMSE ratio results:												
median	.967	.974	.937	.976	.953	.934	.938	.948	.868	.954	.900	.862
prob.<1	.867	.808	.871	.780	.829	.847	.779	.727	.761	.700	.728	.754
5%ile	.899	.897	.787	.901	.811	.770	.765	.754	.583	.757	.582	.545
95%ile	1.011	1.018	1.044	1.022	1.056	1.058	1.073	1.100	1.274	1.119	1.263	1.314
full-sample t-test	.976						.957					
	HP Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.849	.764	.848	.695	.804	.824	.731	.647	.670	.551	.667	.643
MSE-t	.776	.635	.643	.513	.557	.543	.632	.492	.417	.370	.394	.353
ENC-F	.921	.849	.993	.772	.955	.986	.880	.784	.961	.674	.909	.944
ENC-t	.846	.708	.952	.630	.794	.918	.798	.668	.851	.531	.670	.773
RMSE ratio results:												
median	.965	.968	.937	.972	.947	.931	.945	.948	.914	.953	.902	.915
prob.<1	.884	.827	.878	.799	.858	.846	.771	.738	.717	.712	.748	.693
5%ile	.899	.898	.835	.897	.859	.810	.810	.796	.703	.788	.699	.664
95%ile	1.012	1.017	1.035	1.024	1.040	1.065	1.069	1.086	1.198	1.100	1.209	1.276
full-sample t-test	.988						.963					

Notes:
1. The data generating process is based on empirical estimates of equation (1) for the forecast horizon $\tau = 1$ and a simple AR model for the output gap, allowing breaks in the coefficients on lagged inflation in the inflation equation and in the residual variance–covariance matrix. The break in the inflation coefficients occurs in 1980:2. The residual variance of the inflation equation has breaks in 1974:1, 1984:1, and 1991:4; the variance of the gap equation has a break in 1984:1. The inflation equations are given in Table 8. In each of these "true" models, the output gap has predictive power for inflation, in the sense of having a significant regression coefficient.
2. See notes 2-4 of Table 9.

Table 11: Estimates of Models with Breaks in All Slope Coefficients

Dependent variable: $\Delta\pi_t$

Sample: 1958:3 to 2001:4

	GDP price index		**Core CPI**	
	coef. (s.e.)		*coef. (s.e.)*	
	CBO Gap	*HP Gap*	*CBO Gap*	*HP Gap*
constant	.021 (.074)	-.003 (.071)	-.002 (.088)	-.039 (.094)
$\Delta\pi_{t-1}$, $t \leq BP_\pi$	-.260 (.118)	-.293 (.114)	-.111 (.145)	-.187 (.142)
$\Delta\pi_{t-2}$, $t \leq BP_\pi$	-.260 (.090)	-.298 (.090)	-.043 (.140)	-.126 (.131)
$\Delta\pi_{t-1}$, $t > BP_\pi$	-.548 (.097)	-.520 (.100)	-.662 (.142)	-.676 (.141)
$\Delta\pi_{t-2}$, $t > BP_\pi$	-.370 (.109)	-.336 (.106)	-.540 (.154)	-.532 (.145)
gap_{t-1}, $t \leq BP_y$	.128 (.031)	.328 (.085)	.165 (.055)	.359 (.080)
gap_{t-1}, $t > BP_y$	.096 (.113)	.119 (.057)	.045 (.044)	.005 (.122)
$\bar{R}^2$	.154	.187	.259	.285

Notes:

1. The table reports estimates of equation (1) allowing a single break in the set of coefficients on lagged inflation and a single break in the coefficient on the output gap. The inflation coefficient break date BP_π is specified as 1983:1 for the GDP price index and 1980:2 for the core CPI. The output gap coefficient break date BP_y is specified as: 1994:1 for models including the CBO output gap; 1976:2 for the model relating the GDP price index to the HP output gap; and 1983:1 for the model relating core CPI inflation to the HP gap.
2. The reported standard errors are robust with respect to heteroskedasticity.

Table 12: Simulated Power Results for the GDP Price Index with Breaks in All Slope Coefs.

	CBO Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.876	.861	.565	.797	.760	.469	.804	.791	.495	.733	.671	.415
MSE-t	.820	.765	.409	.610	.557	.272	.715	.644	.314	.503	.442	.246
ENC-F	.968	.954	.791	.917	.868	.691	.947	.922	.724	.865	.815	.628
ENC-t	.948	.911	.583	.815	.699	.435	.903	.821	.532	.687	.630	.420
RMSE ratio results:												
median	.967	.952	.987	.943	.966	.992	.916	.873	.975	.859	.916	.993
prob.<1	.902	.900	.681	.859	.856	.611	.846	.846	.599	.818	.775	.523
5%ile	.905	.858	.929	.825	.889	.931	.743	.648	.789	.602	.681	.787
95%ile	1.007	1.015	1.038	1.035	1.022	1.062	1.070	1.074	1.181	1.139	1.127	1.237
full-sample t-test	.983						.961					
	HP Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.590	.436	.666	.355	.649	.622	.404	.355	.520	.329	.489	.484
MSE-t	.453	.254	.441	.152	.398	.368	.297	.203	.312	.133	.265	.275
ENC-F	.962	.917	.882	.869	.861	.821	.914	.854	.781	.765	.754	.742
ENC-t	.902	.775	.701	.623	.625	.580	.791	.613	.583	.478	.522	.494
RMSE ratio results:												
median	.990	1.000	.978	1.014	.971	.978	1.010	1.034	.972	1.057	.970	.975
prob.<1	.634	.498	.740	.423	.750	.694	.463	.420	.627	.402	.595	.612
5%ile	.929	.908	.919	.883	.905	.910	.853	.788	.832	.737	.776	.811
95%ile	1.049	1.101	1.044	1.191	1.046	1.058	1.193	1.329	1.160	1.489	1.219	1.181
full-sample t-test	.998						.989					

Notes:
1. The data generating process is based on empirical estimates of equation (1) for the forecast horizon $\tau = 1$ and a simple AR model for the output gap, allowing breaks in the coefficients on lagged inflation in the inflation equation, the output gap, and the residual variance–covariance matrix. The break in the inflation coefficients occurs in 1983:1. The break in the output gap coefficient occurs in 1994:1 in the case of the CBO gap and in 1976:2 in the case of the HP gap. The residual variance of the inflation equation has breaks in 1970:2 and 1976:2; the variance of the gap equation has a break in 1984:1. The inflation equations are given in Table 11.

2. For each artificial data set, forecasts of $\pi_{t+\tau}^{(\tau)} - \pi_t$ are formed recursively using estimates of equations (1) and (2). These forecasts are then used to form the indicated test statistics, defined in Section 3.2, as well as the ratio of the RMSE of the model including the output gap to the RMSE of the AR model. In each Monte Carlo replication, the simulated test statistics are compared against bootstrapped null critical values (with a significance level of 10%), computed using the procedure described in section 3.3.

3. For the test statistics of interest, the table reports the percentage of Monte Carlo simulations in which the null of equal MSE or forecast encompassing is rejected. For the RMSE ratio, the table reports several summary statistics for the distribution of the ratio in the Monte Carlo trials: the median; the percent of the Monte Carlo ratios that are less than 1 (labeled *prob.<1*); and a 90% interval (given by the *5%ile* and *95%ile* figures). *Full-sample t-test* refers to the usual t–test of the null that the output gap coefficient is less than or equal to zero.

4. The number of Monte Carlo simulations is 1000; the number of bootstrap draws is 2000.

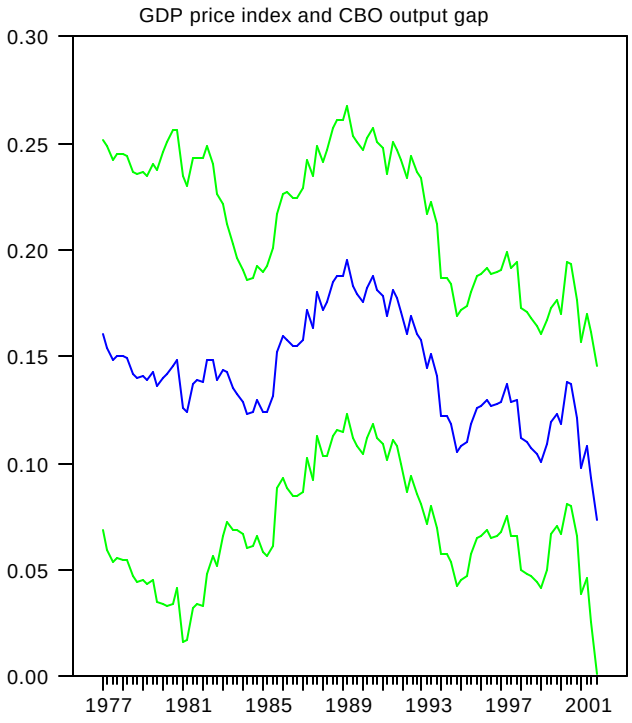
Table 13: Simulated Power Results for the Core CPI with Breaks in All Slope Coefs.

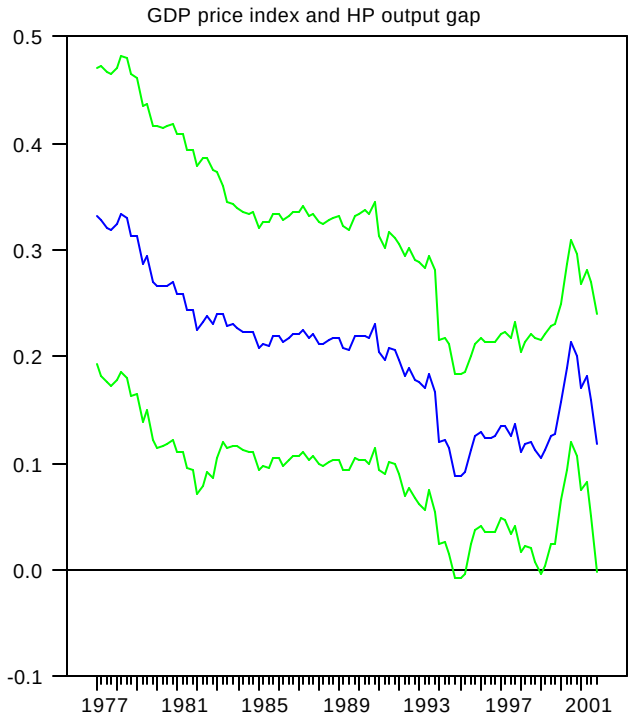
	CBO Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.750	.682	.491	.598	.727	.369	.662	.602	.417	.494	.656	.337
MSE-t	.669	.572	.301	.457	.498	.195	.563	.448	.268	.345	.420	.193
ENC-F	.846	.737	.854	.646	.900	.778	.825	.711	.811	.598	.884	.715
ENC-t	.736	.635	.662	.545	.729	.488	.733	.619	.606	.510	.655	.488
RMSE ratio results:												
median	.977	.977	.993	.978	.957	1.013	.958	.953	1.018	.959	.906	1.076
prob.<1	.806	.782	.540	.750	.807	.417	.713	.714	.462	.679	.713	.391
5%ile	.911	.902	.859	.906	.817	.871	.795	.763	.698	.762	.592	.686
95%ile	1.017	1.020	1.167	1.024	1.054	1.249	1.087	1.100	1.577	1.119	1.277	1.800
full-sample t-test	.958						.931					
	HP Gap											
	$\tau = 1$						$\tau = 4$					
	77-01	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*	*77-01*	*77-89*	*90-01*	*77-84*	*85-92*	*93-01*
MSE-F	.823	.805	.352	.774	.480	.350	.645	.662	.204	.639	.357	.211
MSE-t	.729	.653	.152	.580	.198	.126	.525	.453	.095	.390	.136	.088
ENC-F	.963	.930	.947	.902	.904	.921	.931	.878	.864	.828	.847	.829
ENC-t	.884	.811	.714	.752	.615	.628	.832	.745	.603	.644	.512	.513
RMSE ratio results:												
median	.961	.954	1.018	.952	.991	1.021	.953	.935	1.139	.927	1.047	1.150
prob.<1	.850	.852	.387	.847	.555	.401	.690	.732	.242	.741	.420	.255
5%ile	.886	.874	.914	.863	.887	.902	.805	.767	.837	.739	.758	.816
95%ile	1.018	1.023	1.190	1.025	1.163	1.230	1.109	1.113	1.576	1.117	1.554	1.688
full-sample t-test	.998						.987					

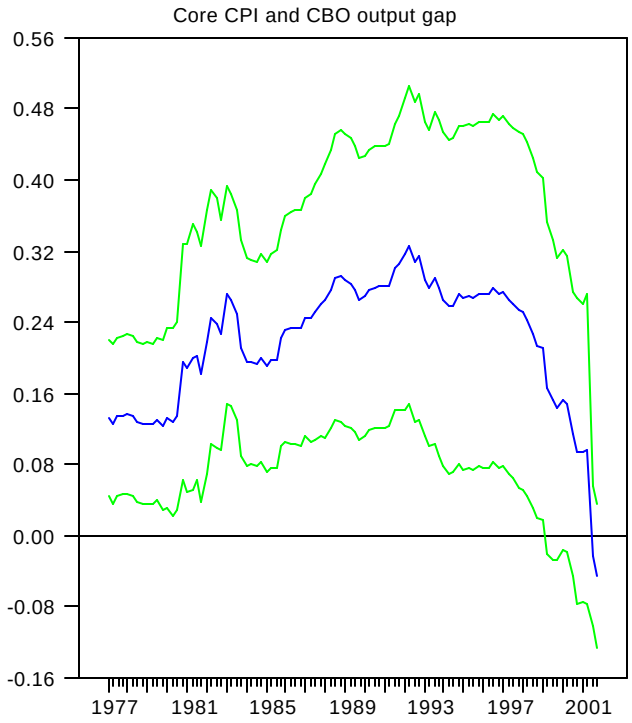
Notes:
1. The data generating process is based on empirical estimates of equation (1) for the forecast horizon $\tau = 1$ and a simple AR model for the output gap, allowing breaks in the coefficients on lagged inflation in the inflation equation, the output gap, and the residual variance–covariance matrix. The break in the inflation coefficients occurs in 1980:2. The break in the output gap coefficient occurs in 1994:1 in the case of the CBO gap and 1983:1 in the case of the HP gap. The residual variance of the inflation equation has breaks in 1974:1, 1984:1, and 1991:4; the variance of the gap equation has a break in 1984:1. The inflation equations are given in Table 11. In each of these "true" models, the output gap has predictive power for inflation, in the sense of having a significant regression coefficient.
2. See notes 2-4 of Table 12.

Figure 1: Rolling Estimates of the Output Gap Coefficient

(coef. est. and 90% confidence interval)







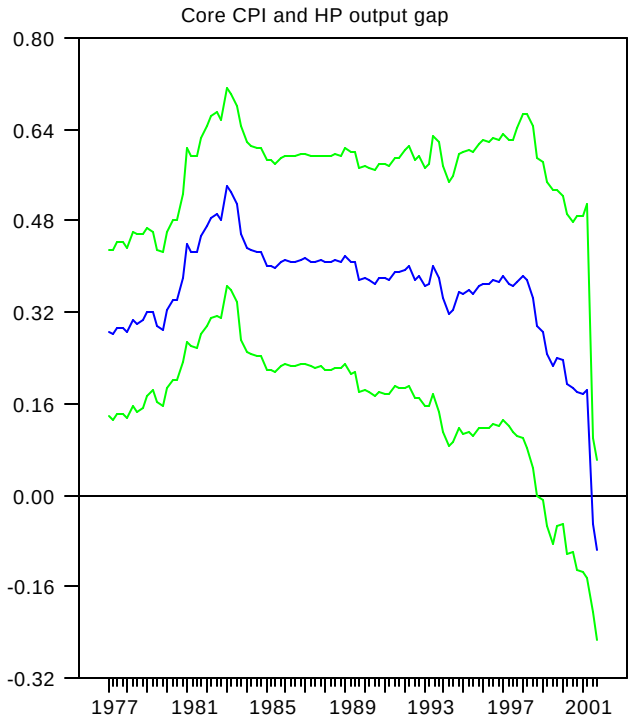


Figure 2: Squared Residuals from Phillips Curve Models

(3 quarter centered moving average)