ISSN 1936-5330

THE FEDERAL RESERVE BANK *of* KANSAS CITY
ECONOMIC RESEARCH DEPARTMENT

Tests of Equal Predictive Ability with Real-Time Data

Todd E. Clark and Michael W. McCracken

July 2007

RWP 07-06

RESEARCH WORKING PAPERS

Tests of Equal Predictive Ability with Real-Time Data *

Todd E. Clark
Federal Reserve Bank of Kansas City

Michael W. McCracken
Board of Governors of the Federal Reserve System

July 2007

Abstract

This paper examines the asymptotic and finite-sample properties of tests of equal forecast accuracy applied to direct, multi–step predictions from both non-nested and nested linear regression models. In contrast to earlier work — including West (1996), Clark and McCracken (2001, 2005), and McCracken (2006) — our asymptotics take account of the real-time, revised nature of the data. Monte Carlo simulations indicate that our asymptotic approximations yield reasonable size and power properties in most circumstances. The paper concludes with an examination of the real–time predictive content of various measures of economic activity for inflation.

JEL Nos.: C53, C12, C52
Keywords: Prediction, real-time data, causality

* *Clark (corresponding author)*: Economic Research Dept.; Federal Reserve Bank of Kansas City; 925 Grand; Kansas City, MO 64198; todd.e.clark@kc.frb.org. *McCracken*: Board of Governors of the Federal Reserve System; 20th and Constitution N.W.; Mail Stop #61; Washington, D.C. 20551; michael.w.mccracken@frb.gov. We gratefully acknowledge helpful comments from Borağan Aruoba, seminar participants at Oregon, SUNY-Albany, and UBC, and participants in the following conferences: Real-Time Data Analysis and Methods at the Federal Reserve Bank of Philadelphia, Computing in Economics in Finance, International Symposium on Forecasting, and NBER Summer Institute. Barbari Rossi provided especially helpful comments in discussing the paper at the Philadelphia Fed conference. 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, Board of Governors, Federal Reserve System, or any of its staff.

1 Introduction

Testing for equal out-of-sample predictive ability is a now common method for evaluating whether a new predictive model forecasts significantly better than an existing baseline model. Various methods have been developed to test whether any gains from the new model are statistically significant. As with in-sample comparisons (e.g. Vuong, 1989), the asymptotic distributions of the test statistics depend on whether the comparisons are between nested or non-nested models. For non-nested comparisons, Granger and Newbold (1977) and Diebold and Mariano (1995) develop asymptotically standard normal tests for predictive ability that allow comparisons between models that don't have estimated parameters. West (1996), McCracken (2000), and Corradi, Swanson and Olivetti (2001) extend these results for non-nested models to allow for estimated parameters; the tests generally continue to be asymptotically standard normal.[1] For nested models, Clark and McCracken (2001, 2005), McCracken (2006), Chao, Corradi and Swanson (2001), and Corradi and Swanson (2002) derive asymptotics for a collection of tests designed to determine whether a nested model forecasts as accurately or encompasses a larger, nesting, model. In most cases, nested comparisons imply asymptotic distributions that are not asymptotically standard normal.[2]

One issue that is uniformly overlooked is the real-time nature of the data. Specifically, the literature ignores the possibility that at any given forecast origin the most recent data is subject to revision. To see how this may be an issue suppose an out-of-sample test of predictive ability is being constructed. The test statistic is functionally very different from an in-sample one and in a fashion that makes it particularly susceptible to changes in the correlation structure of the data as the revision process unfolds. This occurs for three reasons: (i) while parameter estimates are typically functions of only a small number of observations that remain subject to revision, out-of-sample statistics are themselves functions of a sequence of these parameter estimates (one for each forecast origin $t = R, ...T$,), (ii) the predictand used to generate the forecast and (iii) the dependent variable used to construct the forecast error may be subject to revision and hence a sequence of revisions contribute to the test statistic. If it is the case, as noted in Aruoba (2006), that data sub-

[1] However, normality can break down in certain situations, as in the high persistence case examined in Rossi (2005).

[2] Under an alternative asymptotic approximation that treats the estimation sample as fixed (as in a rolling forecasting scheme) rather than limiting to infinity, Giacomini and White (2006) obtain asymptotic normality for a test of equal predictive ability.

ject to revision possess a different mean and covariance structure than final revised data, it is not surprising that tests of predictive ability using real-time data may have a different asymptotic distribution than tests constructed using data that is never revised.

Accordingly, in this paper we provide analytical, Monte Carlo and empirical evidence on pairwise tests of equal out-of-sample predictive ability for models estimated — and forecasts evaluated — using real-time data. We consider comparisons whereby the models are non-nested or nested. In each case we restrict attention to linear direct multi-step (DMS) models evaluated under quadratic loss but do not require that the models be correctly specified; model residuals and forecast errors are allowed to be conditionally heteroskedastic and serially correlated of an order greater than the forecast horizon. We also restrict attention to the case in which parameter estimates are generated on a recursive basis, with the model estimation sample growing as forecasting moves forward in time. Results for the fixed and rolling estimation schemes will be qualitatively similar. As to data revisions, in some cases, we permit the revision process to consist of both "news" and "noise" as defined in Mankiw, Runkle and Shapiro (1984) and applied more recently by Aruoba (2006). In general, though, we emphasize the role of noisy revisions.

Our results indicate substantial differences in the asymptotic behavior of tests of equal predictive ability, relative to those found in the existing literature, when data is subject to revision. For example, when constructing tests of equal predictive ability between non-nested models, West (1996) notes that the effect of parameter estimation error on the test statistic can be ignored when the same loss function is used for estimation and evaluation. In the presence of data revisions, this result continues to hold only in the special case in which the revision process consists only of news. When even some noise is present, parameter estimation error contributes to the asymptotic variance of the test statistic and cannot be ignored when conducting inference.

As another example, when constructing tests of equal predictive ability between nested models, Clark and McCracken (2001, 2005) and McCracken (2006) note that standard test statistics are not asymptotically normal but instead have representations as functions of stochastic integrals. However, when the revision process contains a noise component, we show that the standard test statistics fail not only to be asymptotically normal, but in fact diverge with probability one under the null hypothesis. To avoid this, we introduce a variant of the standard test statistic that is asymptotically standard normal despite being

a comparison between two, recursively estimated, nested models.

Not surprisingly, as with all theoretical results, our conclusions rely upon assumptions made on the observables. What makes our problem specifically troublesome is that the observables are learned sequentially in time across a finite-lived revision process. For any given historical date, we therefore have multiple "observables" for a given dependent or predictor variable. To keep our analytics as transparent as possible, while still remaining relevant for application, we assume that for each variable the revision process continues sequentially for a finite $0 \leq r << R$ periods.[3] We also abstract from other forms of revisions, including benchmark revisions.

While our results are related to the existing literature on tests of out-of-sample predictability, our results also relate back to a literature on forecasting in the presence of data revisions including Howrey (1978), Swanson (1996) and Robertson and Tallman (1998). Notably, our results bear some resemblance to those in Koenig, Dolmas and Piger (2003). They, too, note that the observables likely have different statistical properties depending upon where the observables are in the revision process. They suggest that one can improve forecast accuracy by using the various vintages of data as they would have been observed in real-time to construct forecasts rather than only using those observables that exist in the most recent vintage. Their results differ from ours in that they are interested in forecast accuracy while we are interested in out-of-sample inference but the main issue remains the same: ignoring the data revision process can lead to undesired outcomes — either less accurate forecasts or, in our case, asymptotically invalid inference.

The remainder of the paper proceeds as follows. Section 2 introduces the notation, the forecasting and testing setup, and the assumptions underlying our theoretical results. Section 3 defines the forecast tests considered, provides the null asymptotic results, and lays out how, in practice, asymptotically valid tests can be calculated. Proofs of the asymptotic results are provided in the appendix. Section 4 presents Monte Carlo results on the finite–sample performance of the asymptotics. Section 5 applies our tests to determine whether measures of output have predictive content for U.S. inflation. Section 6 concludes.

[3]For a number of U.S. macroeconomic time series, this simplifying assumption is realistic: e.g., payroll employment and the Conference Board's coincident indicator are subject to a finite number of revisions.

2 Setup

As noted above, in our theory we allow the observables to be subject to revision over a finite number of periods, r. We have in mind the case where r is small relative to the number of observations being used to estimate the model parameters at any given forecast origin. To keep track of the various vintages of a given observation we use the notation $y_s(t)$ to denote the value of the time t vintage of the observation s realization of y. For example, $y_{2000:Q1}(2001{:}Q1)$ refers to the value of y in 2000:Q1 published in the 2001:Q1 vintage. Throughout, when either there is no revision process (so that $r = 0$) or when the revision process is completed (so that $t \geq s + r$), we will drop the notation indexing the vintage and simply let $y_s(t) = y_s$.

The sample of observations $\{\{y_s(t), x_s'(t)\}_{s=1}^{t}\}_{t=R}^{\overline{T}}$ includes a scalar random variable $y_s(t)$ to be predicted, as well as a $(k \times 1)$ vector of predictors $x_s(t)$. When the two models $i = 1, 2$ are nested we let $x_s(t) = x_{2,s}(t) = (x_{1,s}'(t), x_{22,s}'(t))'$ with $x_{i,s}(t)$ the $(k_i \times 1)$ vector of predictors associated with model i. Hence the putatively nested and nesting models are linear regressions with predictors $x_{1,s}(t)$ and $x_{2,s}(t)$ respectively. When the models are non-nested we define $x_{1,s}(t)$ and $x_{2,s}(t)$ as two distinct $(k_i \times 1)$ subvectors of $x_s(t)$ (perhaps having some variables in common).

For each forecast origin t the variable to be predicted is $y_{t+\tau}(t')$, where τ denotes the forecast horizon and $t' \geq t+\tau$ denotes the vintage used to evaluate the forecasts. Throughout the evaluation period, we keep the vintage horizon $r' = t' - t - \tau$ fixed. At the initial forecast origin $t = R$, the present data vintage consists of observations (on $y_s(R)$ and $x_s(R)$) spanning $s = 1, ...R$. Letting $P - \tau + 1$ denote the number of τ–step ahead predictions, the progression of forecast origins span R through $T = R + P - \tau + 1$, each consisting of observations (on $y_s(t)$ and $x_s(t)$) spanning $s = 1, ...t$. The total number of observations in the sample corresponding to the final vintage is $\overline{T} = T + \tau + r'$. Note that the final $\tau + r'$ vintages are used exclusively for evaluation.

Forecasts of $y_{t+\tau}(t')$, $t = R, \ldots, T$, are generated using the two linear models $y_{s+\tau}(t) = x_{1,s}'(t)\beta_1^* + u_{1,s+\tau}(t)$ (model 1) and $y_{s+\tau}(t) = x_{2,s}'(t)\beta_2^* + u_{2,s+\tau}(t)$ (model 2) for $s = 1, ..., t-\tau$. Under the null hypothesis of equal forecast accuracy between nested models, model 2 nests model 1 for all t such that model 2 includes $\dim(x_{22,s}(t)) = k_{22}$ excess parameters. Then $\beta_2^* = (\beta_1^{*\prime}, 0')'$, and $y_{t+\tau}(t') - x_{1,t}'(t)\beta_1^* = u_{1,t+\tau}(t') = u_{2,t+\tau}(t') \equiv u_{t+\tau}(t')$ for all t and t'.

Both β_1^* and β_2^* are re-estimated as we progress across the vintages of data associated

with each forecast origin: for $t = R, \ldots, T$, model i's ($i = 1, 2$) prediction of $y_{t+\tau}(t')$ is created using the parameter estimate $\hat{\beta}_{i,t}$ based on vintage t data. Models 1 and 2 yield two sequences of $P - \tau + 1$ forecast errors, denoted $\hat{u}_{1,t+\tau}(t') = y_{t+\tau}(t') - x'_{1,t}(t)\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+\tau}(t') = y_{t+\tau}(t') - x'_{2,t}(t)\hat{\beta}_{2,t}$, respectively.

Finally, the asymptotic results below use the following additional notation. Let $h_{i,t+\tau}(t') = (y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*)x_{i,t}(t)$, $h_{i,s+\tau} = (y_{s+\tau} - x'_{i,s}\beta_i^*)x_{i,s}$, $H_i(t) = t^{-1}\sum_{s=1}^{t-\tau} h_{i,s+\tau}$, $B_i = (Ex_{i,s}x'_{i,s})^{-1}$ and $d_{t+\tau}(t') = u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t')$. Throughout, when the models are non-nested we let $h_{t+\tau} = (h'_{1,t+\tau}, h'_{2,t+\tau})'$, $h_{t+\tau}(t') = (h'_{1,t+\tau}(t'), h'_{2,t+\tau}(t'))'$ and $U_{t+\tau} = [d_{t+\tau}(t'), h'_{t+\tau}(t') - Eh'_{t+\tau}(t'), h'_{t+\tau}, x'_t - Ex'_t]'$. When the models are nested, let $h_{s+\tau} = h_{2,s+\tau}$, $h_{t+\tau}(t') = h_{2,t+\tau}(t')$ and $U_{t+\tau} = [h'_{t+\tau}(t') - Eh'_{t+\tau}(t'), h'_{t+\tau}, x'_t - Ex'_t]'$.[4] In either case let $H(t) = t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau}$. Define the selection matrix $J = (I_{k_1 \times k_1}, 0_{k_1 \times k_{22}})$ and let Ω denote the asymptotic variance of the scaled average loss differential defined more precisely in Section 3.

Given the definitions and forecasting scheme described above, the following assumptions are used to derive the limiting distributions in Theorems 1-4. The assumptions are intended to be only sufficient, not necessary and sufficient.

(A1) The parameter estimates $\hat{\beta}_{i,t}$, $i = 1, 2$, $t = R, ..., T$, are estimated using OLS for each vintage in succession and hence satisfy $\hat{\beta}_{i,t} = \arg\min_{\beta_i} \sum_{s=1}^{t-\tau}(y_{s+\tau}(t) - x'_{i,s}(t)\beta_i)^2$.

(A2) (a) $U_{t+\tau}$ is covariance stationary, (b) $EU_{t+\tau} = 0$, (c) $Ex_tx'_t < \infty$ and is positive definite, (d) For some $n > 1$ and for each integer $0 \leq j$, $(y_t(t+j), x'_t(t+j))'$ is uniformly L^{4n} bounded, (e) $U_{t+\tau}$ is strong mixing with coefficients of size $-4n/(n-1)$, (f) Ω is positive definite.

(A3) (a) Let $K(x)$ be a kernel such that for all real scalars x, $|K(x)| \leq 1$, $K(x) = K(-x)$ and $K(0) = 1$, $K(x)$ is continuous, and $\int_{-\infty}^{\infty} |K(x)|dx$, (b) For some bandwidth M and constant $m \in (0, 0.5)$, $M = O(P^m)$.

(A4) $\lim_{R,P\to\infty} P/R = \pi \in (0, \infty)$.

(A4′) $\lim_{R,P\to\infty} P/R = 0$.

The assumptions provided here are closely related to those in West (1996) and Clark

[4] When the models are nested, $d_{t+\tau}(t') = 0$ for all t and t'.

and McCracken (2005). We restrict attention to forecasts generated using parameters estimated by OLS (Assumption 1) and while we do not allow for processes with either unit roots or time trends, we do allow for conditional heteroskedasticity and serial correlation in the levels and squares of the forecast errors (Assumption 2). When long-run variances are estimated, standard kernel estimators are used (Assumption 3). We provide asymptotic results for situations in which the in-sample size of the initial forecast origin R and the number of predictions P are of the same order (Assumption 4) as well as when R is large relative to P (Assumption 4′).

3 Tests and Asymptotic Distributions

In this section we provide asymptotics for tests of equal forecast accuracy for non-nested and nested comparisons. For the comparison of non-nested models we allow data revisions to consist of both news and noise. In the nested case, for tractability we focus on data revisions consisting only of noise, but discuss the impact of news-only revisions.

3.1 Non-nested comparisons

In the context of non-nested models, Diebold and Mariano (1995) propose a test for equal MSE based upon the sequence of loss differentials $\hat{d}_{t+\tau}(t') = \hat{u}^2_{1,t+\tau}(t') - \hat{u}^2_{2,t+\tau}(t')$. If we define $\text{MSE}_i = (P-\tau+1)^{-1}\sum_{t=R}^{T}\hat{u}^2_{i,t+\tau}(t')$ $(i=1,2)$, $\bar{d} = (P-\tau+1)^{-1}\sum_{t=R}^{T}\hat{d}_{t+\tau}(t') = \text{MSE}_1 - \text{MSE}_2$, $\hat{\Gamma}_{dd}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T}(\hat{d}_{t+\tau}(t')-\bar{d})(\hat{d}_{t+\tau-j}(t'-j)-\bar{d})$, $\hat{\Gamma}_{dd}(-j) = \hat{\Gamma}_{dd}(j)$, and $\hat{S}_{dd} = \sum_{j=-P+1}^{P-1}K(j/M)\hat{\Gamma}_{dd}(j)$, the statistic takes the form

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

Under the null that the population difference in MSEs from models 1 and 2 equal zero, the authors argue that the test statistic is asymptotically standard normal and hence inference can be conducted using the relevant tables.

West (1996), however, notes that this outcome depends upon whether or not the forecast errors depend upon estimated parameters. Specifically, if linear, OLS-estimated models are used for forecasting, then $P^{1/2}\bar{d} \to^d N(0,\Omega)$, where $\Omega = S_{dd} + 2(1-\pi^{-1}\ln(1+\pi))(FBS_{dh} + FBS_{hh}BF')$ with $F = (-2Eu_{1,t+\tau}x'_{1,t}, 2Eu_{2,t+\tau}x'_{2,t})$, B a block diagonal matrix with block diagonal elements B_1 and B_2, S_{dd} the long-run variance of $d_{t+\tau}$, S_{hh} the long-run variance of $h_{t+\tau}$ and S_{dh} the long-run covariance of $h_{t+\tau}$ and $d_{t+\tau}$. As a result, the MSE-t test as

constructed in (1) may be missized because, generally speaking, the estimated variance $\hat{S}_{dd}$ is consistent for S_{dd} but not Ω.

One case in which the MSE-t test (1) will be asymptotically valid in the presence of estimated parameters is when $F = 0$. This case arises naturally in the present context because F is equal to zero when the forecast error is uncorrelated with the predictors — a case that will hold when quadratic loss is used for both estimation and inference on predictive ability and the observables are covariance stationary. However, when the data is subject to revision, the population level residuals $y_{s+\tau} - x'_{i,s}\beta_i^*$, $s = 1, ..., t-\tau$, and forecast errors $y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*$, $t = R, ..., T$, need not have the same covariance structure. Consequently, $E(y_{s+\tau} - x'_{i,s}\beta_i^*)x_{i,s}$ equaling zero need not imply anything about whether or not $E(y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*)x_{i,t}(t)$ equals zero.

If we keep track of this distinction, we obtain the following expansion.

Lemma 1: Let Assumptions 1, 2 and 4 or 4′ hold. $P^{1/2}\bar{d} = P^{-1/2}\sum_{t=R}^{T}(u_{1,t+\tau}^2(t') - u_{2,t+\tau}^2(t') + FBH(t)) + o_p(1)$.

The expansion in Lemma 1 is notationally identical to that in West's (1996) Lemma 4.1. Conceptually, though, it differs in two important ways. First, the analytics are derived allowing for data revisions at the end of each sequential vintage of data. Second, F is defined as $2(-Eu_{1,t+\tau}(t')x'_{1,t}(t), Eu_{2,t+\tau}(t')x'_{2,t}(t))$, thus emphasizing the distinction between the population in-sample residuals and the population out-of-sample forecast errors. Since the asymptotic expansion is notationally identical to West's (1996), the asymptotic distribution of the scaled average of the loss differentials remains (notationally) the same.

Theorem 1: Let Assumptions 1, 2 and 4 or 4′ hold. $P^{1/2}\bar{d} \rightarrow^d N(0, \Omega)$ where $\Omega = S_{dd} + 2(1 - \pi^{-1}\ln(1+\pi))(FBS_{dh} + FBS_{hh}BF')$.

Since the asymptotic distribution is essentially the same as in West (1996), the special cases in which one can ignore parameter estimation error remain essentially the same. For example, if the number of forecasts $P - \tau + 1$ is small relative to the number of in-sample observations from the initial forecast origin R, such that $\lim_{R,P\rightarrow\infty} P/R = \pi = 0$, then $2(1 - \pi^{-1}\ln(1+\pi)) = 0$, and hence the latter covariance terms are zero. This case is identical to that in West (1996).

Another special case arises when the out-of sample moment condition $F = 2(-Eu_{1,t+\tau}(t')x'_{1,t}(t), Eu_{2,t+\tau}(t')x'_{2,t}(t))$ equals zero. In this case the latter covariance terms are zero and hence parameter estimation error can be ignored. To see when this will or will

not arise it is useful to write out the population forecast errors explicitly. That is, consider the moment condition $E(y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*)x'_{i,t}(t)$. Moreover, note that β_i^* is defined as the probability limit of the regression parameter estimate in the regression $y_{s+\tau} = x'_{i,s}\beta_i^* + u_{i,s+\tau}$. Hence F equals zero if $Ex_{i,t}(t)y_{t+\tau}(t') = (Ex_{i,t}(t)x'_{i,t}(t))(Ex_{i,t}x'_{i,t})^{-1}(Ex_{i,t}y_{t+\tau})$ for each $i = 1, 2$. Some specific instances that result in $F = 0$ are listed below.

1. x and y are unrevised
2. x is unrevised and the revisions to y are uncorrelated with x
3. x is unrevised and final revised vintage y is used for evaluation
4. x is unrevised and the "vintages" of y's are redefined so that the data release used for estimation is also used for evaluation (as suggested by Koenig, Dolmas and Piger (2001))

In general though, neither of these special cases — that $\pi = 0$ or $F = 0$ — need hold. In the former case, West and McCracken (1998) emphasize that in finite samples the ratio $P/R = \widehat{\pi}$ may be small but that need not guarantee that parameter estimation error is negligible since it may be the case that $FBS_{dh} + FBS_{hh}BF'$ remains large. For the latter case, in the presence of predictable data revisions it is typically not the case that $F = 0$. To conduct inference then requires constructing a consistent estimate of the asymptotic variance Ω given in Theorem 1. We return to consistent estimation of Ω in Section 3.3.

To illustrate exactly how real time revisions can create a non–zero covariance between real-time forecast errors and predictors, consider a simple DGP in which the final data (for t), published with a one-period delay (in $t+1$), are generated by

$$\begin{aligned} y_t &= \beta x_{1,t-1} + \beta x_{2,t-1} + e_{y,t} + v_{y,t} \\ x_{i,t} &= e_{x_i,t} + v_{x_i,t},\ i = 1, 2 \\ & \quad e_{y,t}, v_{y,t}, e_{x_1,t}, v_{x_1,t}, e_{x_2,t}, v_{x_2,t}\ iid\ N(0, \cdot), \end{aligned} \tag{2}$$

where e's represent innovation components that will be included in initial estimates, v's represent news components that will not, $\text{var}(e_{x_i,t}) = \sigma^2_{e,x}$, and $\text{var}(v_{x_i,t}) = \sigma^2_{v,x}$ for $i = 1, 2$. Initial estimates for period t, published in t and denoted $y_t(t)$, $x_{1,t}(t)$, and $x_{2,t}(t)$, contain news and noise:

$$\begin{aligned} y_t(t) &= y_t - v_{y,t} + w_{y,t} \\ x_{i,t}(t) &= x_{i,t} - v_{x_i,t} + w_{x_i,t},\ i = 1, 2 \\ & \quad w_{y,t}, w_{x_1,t}, w_{x_2,t}\ iid\ N(0, \cdot), \end{aligned} \tag{3}$$

where v's correspond to the news component of revisions, w's denote the noise in the initial

estimates, and $\text{var}(w_{x_i,t}) = \sigma^2_{w,x}$ $i = 1, 2$. Finally, suppose forecasts are generated from two models of the form $y_{t+1} = b_i x_{i,t} + u_{i,t+1}$, $i = 1, 2$.

The population value of the real–time forecast error for model i (generated in period t using data available in t, and evaluated using initial estimates available in $t+1$) is

$$u_{i,t+1}(t+1) = y_{t+1}(t+1) - \beta x_{i,t}(t) = e_{y,t+1} + w_{y,t+1} + \beta x_{j,t} + \beta v_{x_i,t} - \beta w_{x_i,t}. \quad (4)$$

The noise component $w_{x_i,t}$ creates a non–zero covariance between the real time forecast error and predictor, giving rise to a non–zero F matrix. For forecast i, this covariance is

$$\begin{aligned} E\left[u_{i,t+1}(t+1)x_{i,t}(t)\right] &= E\left[\left(e_{y,t+1} + w_{y,t+1} + \beta x_{j,t} + \beta v_{x_i,t} - \beta w_{x_i,t}\right)\left(e_{x_i,t} + w_{x_i,t}\right)\right] \quad (5) \\ &= -\beta\sigma^2_{w,x}. \end{aligned}$$

Nonetheless, in practice, even with $F \neq 0$, it is possible that a negative impact of FBS_{dh} could offset the positive impact of the variance component $FBS_{hh}BF'$. In such a setting, the correction necessitated by predictable data revisions may be small, either positive or negative. At least in the DGPs we consider, it looks like this may often be the case. To see why, consider the simple DGP above. In this case, $F = 2(\beta\sigma^2_{w,x},\ -\beta\sigma^2_{w,x})'$. Letting $\sigma^2_x = \sigma^2_{e,x} + \sigma^2_{v,x}$ and $\sigma^2_u = \sigma^2_{e,y} + \sigma^2_{v,y}$, simple algebra yields

$$S_{hh} = \begin{pmatrix} \sigma^2_u\sigma^2_x + \beta^2\sigma^4_x & \beta^2\sigma^4_x \\ \beta^2\sigma^4_x & \sigma^2_u\sigma^2_x + \beta^2\sigma^4_x \end{pmatrix}, \quad (6)$$

and $S_{dh} = 2\beta\sigma^2_{e,x}\sigma^2_{e,y}[-1, 1]'$. Putting together all the pieces yields a population–level variance correction of

$$FBS_{dh} + FBS_{hh}BF' = \frac{8\beta^2\sigma^2_{w,x}}{\sigma^2_x}\left(\sigma^2_{w,x}\sigma^2_u - \sigma^2_{e,y}\sigma^2_{e,x}\right). \quad (7)$$

As this shows, the positive impact of noise (through $\sigma^2_{w,x}$) on the variance correction can be offset or even dominated by the negative impact associated with the information content in initial releases of y and x_1 and x_2 (through $\sigma^2_{e,y}$ and $\sigma^2_{e,x}$).

3.2 Nested comparisons

In the context of nested models, Clark and McCracken (2005) and McCracken (2006) also propose tests for equal MSE based upon the sequence of loss differentials. Specifically, they consider the MSE-t statistic (1) applied to nested models and another test that can be constructed analogously to an in-sample F-test but using out-of-sample forecast errors:

$$\text{MSE-}F = (P - \tau + 1) \times \frac{\text{MSE}_1 - \text{MSE}_2}{\text{MSE}_2} = (P - \tau + 1) \times \frac{\bar{d}}{\text{MSE}_2}. \quad (8)$$

In both cases, the tests have limiting distributions that are non–standard when the forecasts are nested under the null. Specifically, McCracken (2006) shows that, for one–step ahead forecasts from well-specified nested models, the MSE-t and MSE-F statistics converge in distribution to functions of stochastic integrals of quadratics of Brownian motion, with limiting distributions that depend on the parameter π and the number of exclusion restrictions k_{22}, but not any unknown nuisance parameters. For this case, simulated asymptotic critical values are provided. In Clark and McCracken (2005), the asymptotics are extended to permit direct multi-step forecasts and conditional heteroskedasticity. In this environment the limiting distributions are affected by unknown nuisance parameters. Accordingly, for this situation, a bootstrap procedure is recommended. However, all of these results are derived ignoring the potential for data revisions.

In the presence of predictable data revisions, the asymptotics for tests of predictive ability change dramatically — much more so than in the non-nested case. Again, the issue is that when there are data revisions, the residuals $y_{s+\tau} - x'_{i,s}\beta_i^*$ $s = 1, ..., t-\tau$ and the forecast errors $y_{t+\tau}(t') - x'_{i,t}(t)\beta_i^*$ $t = R, ..., T$ need not have the same covariance structure and hence $F = 2(Eu_{2,t+\tau}(t')x'_{2,t}(t))$ need not equal zero. If we keep track of this distinction, we obtain the following expansion.

Lemma 2: Let Assumptions 1 and 2 hold and let $F(-JB_1J'+B_2) \neq 0$. (i) If Assumption 4 holds, $P^{1/2}\bar{d} = F(-JB_1J' + B_2)(P^{-1/2}\sum_{t=R}^{T} H(t)) + o_p(1)$. (ii) If Assumption 4′ holds, $R^{1/2}\bar{d} = F(-JB_1J' + B_2)(R^{1/2}H(R)) + o_p(1)$.

The expansion in Lemma 2 (i) bears some resemblance to that in Lemma 1 for non-nested models but omits the lead term $(P^{-1/2}\sum_{t=R}^{T} u_{1,t+\tau}^2(t') - u_{2,t+\tau}^2(t'))$ because the models are nested under the null. Interestingly, neither (i) nor (ii) bears any resemblance to the corresponding expansions in Clark and McCracken (2005) and McCracken (2006) for nested models. The key difference is that the Lemma 2 expansion is of order $P^{1/2}$, rather than order P as in Clark and McCracken (2005) and McCracken (2006). Not surprisingly, this change in order implies very different asymptotic behavior of out-of-sample averages of loss differentials from nested models.

Theorem 2: Let Assumptions 1 and 2 hold and let $F(-JB_1J' + B_2) \neq 0$. (i) If Assumption 4 holds, $P^{1/2}\bar{d} \rightarrow^d N(0, \Omega)$, where $\Omega = 2(1 - \pi^{-1}\ln(1+\pi))F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$. (ii) If Assumption 4′ holds, $R^{1/2}\bar{d} \rightarrow^d N(0, \Omega)$, where $\Omega = F(-JB_1J' + B_2)S_{hh}(-JB_1J' + B_2)F'$.

Theorem 2 makes clear that in the presence of predictable revisions, a t-test for equal predictive ability can be constructed that is asymptotically standard normal under the null hypothesis. This is in sharp contrast to the results in Clark and McCracken (2005) and McCracken (2006), in which the tests generally have non-standard limiting distributions. This finding has a number of important implications, listed below.

1. The MSE-t test (1) diverges with probability 1 under the null hypothesis. To see this note that by Theorem 2, the numerator of MSE-t is $O_p(1)$. Following arguments made in Clark and McCracken (2005) and McCracken (2006), the denominator of the MSE-t is $O_p(P^{-1})$. Taking account of the square root in the denominator of the MSE-t test implies that the MSE-t test is $O_p(P^{1/2})$ and hence the MSE-t test has an asymptotic size of 50%. A similar argument implies the MSE-F also diverges.

2. Out-of-sample inference for nested comparisons can be conducted without the strong auxiliary assumptions made in Clark and McCracken (2005) and McCracken (2006) regarding the correct specification of the models.[5]

3. Perhaps most importantly, asymptotically valid inference can be conducted without the bootstrap or non-standard tables. So long as an asymptotically valid estimate of Ω is available, standard normal tables can be used to conduct inference. Consistent methods for estimating the appropriate standard errors are described in Section 3.3.

Regardless, it is possible that the asymptotic distribution of the MSE-t test can differ from that given in Theorem 2. The leading case occurs when the revisions consist of news rather than noise so that $F = 0$.[6] But even with predictable revisions (that make F non-zero), Theorem 2 fails to hold when $F(-JB_1J' + B_2)$ (and hence Ω) equals zero. In both cases we have established that the MSE-t (from (1)) is bounded in probability under the null. However, in each instance the asymptotic distributions are non-standard in much the same way as the results in Clark and McCracken (2005) for nested models. Moreover, conducting inference using these distributions is complicated by the presence of unknown nuisance parameters. We leave a complete characterization of these distributions to future research. In the Monte Carlo section, however, we examine the ability of the distributions developed in this paper and in Clark and McCracken (2005) to reasonably approximate the

[5]In previous work we have required that serial correlation in the residuals and forecast errors were of finite order. In most instances we treated τ-step ahead errors as forming an MA($\tau - 1$) process.

[6]It will also be the case that $F = 0$ in exchange rate forecasting applications in which the null model is a random walk and the alternative includes variables subject to predictable revisions.

more complicated, true asymptotic distributions.

3.3 Estimating the standard errors

To construct asymptotically valid estimates of the above standard errors, some combination of S_{dd}, S_{dh}, S_{hh}, F, B, and $2(1-\pi^{-1}\ln(1+\pi))$ needs to be estimated. Since $\hat{\pi}=P/R$ is consistent for π, estimating $\Pi \equiv 2(1-\pi^{-1}\ln(1+\pi))$ is trivial. For F and B we use the obvious sample analogs. For $\widehat{B}_i = (T^{-1}\sum_{s=1}^{T-\max(\tau,r)} x_{i,s}x_{i,s}')^{-1}$, we let $\widehat{B}$ denote the block diagonal matrix constructed using $\widehat{B}_1$ and $\widehat{B}_2$. For non-nested comparisons, we define $\widehat{F}_i = 2(-1)^i[P^{-1}\sum_{t=R}^{T}\hat{u}_{i,t+\tau}(t')x_{i,t}'(t)]$ and $\widehat{F}=(\widehat{F}_1,\widehat{F}_2)$. For nested comparisons, $\widehat{F} = 2[P^{-1}\sum_{t=R}^{T}\hat{u}_{2,t+\tau}(t')x_{2,t}'(t)]$.

For the long-run variances and covariances we consider estimates based upon standard kernel-based estimators akin to those used in West (1996), West and McCracken (1998) and McCracken (2000). To be more precise, we use kernel-weighted estimates of $\Gamma_{dd}(j) = Ed_{t+\tau}(t')d_{t+\tau-j}(t'-j)$, $\Gamma_{dh}(j) = Ed_{t+\tau}(t')h_{t+\tau-j}'$ and $\Gamma_{hh}(j) = Eh_{t+\tau}h_{t+\tau-j}'$ to estimate S_{dd}, S_{dh} and S_{hh}. To construct the relevant pieces recall that $\hat{u}_{i,t+\tau}(t') = y_{t+\tau}(t') - x_{i,t}'(t)\hat{\beta}_{i,t}$, $t=R,...,T$. For non-nested comparisons, define $\widehat{h}_{s+\tau} = ((y_{s+\tau}-x_{1,s}'\hat{\beta}_{1,T})x_{1,s}', (y_{s+\tau}-x_{2,s}'\hat{\beta}_{2,T})x_{2,s}')'$, $s=1,...,T$. For nested comparisons, define $\widehat{h}_{s+\tau} = (y_{s+\tau}-x_{2,s}'\hat{\beta}_{2,T})x_{2,s}$, $s=1,...,T$.

With these sequences of forecast errors and OLS orthogonality conditions in hand, let $\hat{\Gamma}_{dd}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T}(\hat{d}_{t+\tau}(t')-\bar{d})(\hat{d}_{t+\tau-j}(t'-j)-\bar{d})$, $\hat{\Gamma}_{hh}(j) = T^{-1}\sum_{s=1+j}^{T}\widehat{h}_{s+\tau}\widehat{h}_{s+\tau-j}'$ and $\hat{\Gamma}_{dh}(j) = P^{-1}\sum_{t=R+j}^{T}\hat{d}_{t+\tau}(t')\widehat{h}_{t+\tau-j}'$, with $\hat{\Gamma}_{dd}(j)=\hat{\Gamma}_{dd}(-j)$, $\hat{\Gamma}_{hh}(j)=\hat{\Gamma}_{hh}'(-j)$ and $\hat{\Gamma}_{dh}(j)=\hat{\Gamma}_{dh}'(-j)$. Let $K(\cdot)$ define an appropriate kernel function and M a bandwidth. We then estimate the long-run variances and covariances as $\hat{S}_{dd} = \sum_{j=-P+1}^{P-1}K(j/M)\hat{\Gamma}_{dd}(j)$, $\hat{S}_{hh} = \sum_{j=-T+1}^{T-1}K(j/M)\hat{\Gamma}_{hh}(j)$, and $\hat{S}_{dh} = \sum_{j=-P+1}^{P-1}K(j/M)\hat{\Gamma}_{dh}(j)$. The following theorem shows that the relevant pieces are consistent for their population counterparts.

Theorem 3: Let Assumptions 1, 2 and 4 or 4′ hold. (a) $\widehat{B}_i \rightarrow^p B_i$, $\widehat{F} \rightarrow^p F$, $\hat{\Gamma}_{dd}(j) \rightarrow^p \Gamma_{dd}(j)$, $\hat{\Gamma}_{dh}(j) \rightarrow^p \Gamma_{dh}(j)$ and $\hat{\Gamma}_{hh}(j) \rightarrow^p \Gamma_{hh}(j)$. (b) If Assumption 3 holds, $\hat{S}_{dd} \rightarrow^p S_{dd}$, $\hat{S}_{dh} \rightarrow^p S_{dh}$, $\hat{S}_{hh} \rightarrow^p S_{hh}$.

Along with Theorems 1-2, Theorem 3 and Slutsky's Theorem imply that $P^{1/2}\bar{d}/\widehat{\Omega}^{1/2}$ (or $R^{1/2}\bar{d}/\widehat{\Omega}^{1/2}$) is asymptotically standard normal and hence asymptotically valid inference can be conducted using the appropriate tables.

4 Monte Carlo Evidence

We proceed by first describing our Monte Carlo framework and the construction of the test statistics. We then present results on the size and power of the forecast–based tests, first for the non–nested case and then the nested case. The tests are applied at forecast horizons of one and four steps. The DGPs include simple ones for which we can work out analytic results and more complicated ones parameterized to roughly reflect the properties of the change in the quarterly U.S. inflation rate (as measured by the GDP price index) and an output gap computed with the HP filter. In these cases, the variable being forecast roughly corresponds to the change in inflation; the variables used to forecast inflation have properties similar to those of the HP output gap.

4.1 Monte Carlo design: non-nested case

In the non–nested forecast case, we consider three DGPs patterned broadly after those in Godfrey and Pesaran (1983). For DGPs 1 and 2, the final data are generated by the same basic system (equation (2)) used as an example in section 3.1:

$$\begin{aligned} y_t &= .4x_{1,t-1} + (.4+\beta)x_{2,t-1} + \sigma_{e,y}e_{y,t} + \sigma_{v,y}v_{y,t} \\ x_{i,t} &= \sigma_{e,x}e_{x_i,t} + \sigma_{v,x}v_{x_i,t},\ i = 1,2 \\ & \quad e_{y,t}, v_{y,t}, e_{x_1,t}, v_{x_1,t}, e_{x_2,t}, v_{x_2,t}\ iid\ N(0,1). \end{aligned} \tag{9}$$

Across the DGP 1 and 2 experiments, the variances of the y and x variables are held fixed, but the variances of the innovation components vary, as described below. For DGP 3, the final data are generated by

$$\begin{aligned} y_t &= -.4y_{t-1} - .3y_{t-2} + .25x_{1,t-1} + (.25+\beta)x_{2,t-1} + \sigma_{e,y}e_{y,t} + \sigma_{v,y}v_{y,t} \\ x_{i,t} &= 1.1x_{i,t-1} - .3x_{i,t-2} + \sigma_{e,x}e_{x_i,t} + \sigma_{v,x}v_{x_i,t},\ i = 1,2 \\ & \quad e_{y,t}, v_{y,t}, e_{x_1,t}, v_{x_1,t}, e_{x_2,t}, v_{x_2,t}\ iid\ N(0,1). \end{aligned} \tag{10}$$

For all DGPs, the coefficient β is set to zero in size experiments. In power experiments, β is set to 0.6 in DGPs 1 and 2 and 0.75 in DGP 3.

We focus, of course, on data subject to revision, supposing the final values are released with a delay. In practice, data such as GDP are subject to many revisions. In the case of GDP-related data, three estimates are published 1, 2, and 3 months after the end of a quarter; subsequent estimates are published in three annual revisions; and yet further

revisions are published in periodic benchmark revisions. In our Monte Carlo exercises, we try to simplify matters while at the same time preserving some of the essential features of actual (non-benchmark) revisions. We assume a single revision of an initially published estimate. For analytical tractability, in DGPs 1 and 2 the final values are published with just a one-period delay. In DGP 3, the final values are published with a four-period delay. Specifically, a first estimate of each variable's value in period t is published in period t (denoted $y_t(t)$, $x_{1,t}(t)$, and $x_{2,t}(t)$). The final estimates (y_t, $x_{1,t}$, and $x_{2,t}$) are treated as being published in period $t+1$ in DGPs 1 and 2 and period $t+4$ in DGP 3. While the particular dating is arbitrary, our intention is to capture the empirical regularity of early revisions.

Motivated by work in such studies as Croushore and Stark (2003), Faust and Wright (2005), and Arouba (2006) on predictability in data revisions, the revision processes have a common general structure, relating a revision between the prior estimate and current estimate to the prior estimate and an independent innovation:

$$\begin{aligned} y_t(t) &= y_t - \sigma_{v,y} v_{y,t} + \sigma_{w,y} w_{y,t} \\ x_{i,t}(t) &= x_{i,t} - \sigma_{v,x} v_{x_i,t} + \sigma_{w,x} w_{x_i,t},\ i = 1,2 \\ & \quad w_{y,t}, w_{x_1,t}, w_{x_2,t}\ iid\ N(0,1) \end{aligned} \tag{11}$$

With this structure, the initial estimates include all of the information in the final value except the innovation components denoted by v's (incorporated in, e.g., $y_t - \sigma_{v,y} v_{y,t}$), but add in measurement error components denoted by w's. Revisions (final less initial) are then a linear combination of news (v's) and noise (w's). Setting the variances of the w terms to zero yields revisions that are comprised entirely of news.

For DGPs 2 and 3, our parameterizations of the revision processes are roughly drawn from evidence in Aruoba (2006) and empirical estimates for real-time U.S. data on the change in GDP inflation and the HP output gap from 1965 through 2003.[7] For DGP 1, however, we use a parameterization designed to yield a more sizable impact of data revisions on real time forecast inference. Specifically, in DGP 1 experiments, the innovation variances are set to $\sigma^2_{e,y} = .1$, $\sigma^2_{v,y} = .9$, $\sigma^2_{w,y} = .2$, $\sigma^2_{e,x} = 1.7$, $\sigma^2_{v,x} = .3$, and $\sigma^2_{w,x} = 2$. Under this parameterization, the correlation of the revision in y with the initial estimate is about -0.2,

[7] More specifically, using the first $k = 4$ vintages, we estimate for each variable the correlation of the revision from vintage k to $k+1$ with the vintage k data series. For each variable, we then averaged the resulting three correlations.

in line with our data. However, the revision variance is nearly 70 percent of the variance of y, well above the 30 percent average reported by Aruoba (2006). The correlation of the revision in each x variable with the initial estimate is nearly -0.7, just a tad higher than in actual data for the output gap. But the revision variance of each x variable is 15 percent larger than the variance of the corresponding final series.

In DGP 2 experiments, we use $\sigma^2_{e,y} = .8$, $\sigma^2_{v,y} = .2$, $\sigma^2_{w,y} = .2$, $\sigma^2_{e,x} = 1.7$, $\sigma^2_{v,x} = .3$, and $\sigma^2_{w,x} = .5$. In DGP 2, the correlation of the revision in y with the initial estimate remains around -0.2, roughly in line with our data. The correlation of the revision in the x variables with the initial estimate is nearly -0.4, which is lower than in our actual data on the output gap, but not out of line with evidence for other variables. As a share of the variance of the final data, the variance of revisions is about 20 percent for y and 40 percent for the x variables. These settings balance evidence from our own data with the broader results in Aruoba (2006). Finally,we parameterize DGP 3 to obtain magnitudes of revisions and predictability in line with DGP 2, setting the following: $\sigma^2_{e,y} = .8$, $\sigma^2_{v,y} = .2$, $\sigma^2_{w,y} = .2$, $\sigma^2_{e,x} = .2$, $\sigma^2_{v,x} = .3$, and $\sigma^2_{w,x} = .5$.

With DGPs 1 and 2, we test for equal accuracy of τ-horizon forecasts from models

$$y^{(\tau)}_{t+\tau} = a_1 x_{1,t} + u_{1,t+\tau} \tag{12}$$

$$y^{(\tau)}_{t+\tau} = b_1 x_{2,t} + u_{2,t+\tau}, \tag{13}$$

where $y^{(\tau)}_{t+\tau} \equiv \tau^{-1} \sum_{s=1}^{\tau} y_{t+s}$. The forecasting models for DGP 3 experiments take the form:

$$y^{(\tau)}_{t+\tau} = a_0 + a_1 y_t + a_2 y_{t-1} + a_3 x_{1,t} + u_{1,t+\tau} \tag{14}$$

$$y^{(\tau)}_{t+\tau} = b_0 + b_1 y_t + b_2 y_{t-1} + b_3 x_{2,t} + u_{2,t+\tau}. \tag{15}$$

At each forecast origin t, the observable time series for each variable consists of an initial or first vintage estimates for period $t-r+1$ through t and final values for periods $t-r$ and earlier, where $r = 1$ in DGPs 1 and 2 and $r = 4$ in DGP 3. As forecasting moves forward, the models are recursively re-estimated with an expanding sample of data, by OLS.

In evaluating forecasts, we compute forecast errors using actual values of y taken to be the initial estimate published in period t, $y_t(t)$. We form two versions of the MSE-t test, one with a standard error of just an estimate $\widehat{S}_{dd}$ and the other with an estimate $\widehat{\Omega} = \widehat{S}_{dd} + 2\widehat{\Pi}(\widehat{F}\widehat{B}\widehat{S}_{dh} + \widehat{F}\widehat{B}\widehat{S}_{hh}\widehat{B}\widehat{F}')$. We compute the long-run variances $\widehat{S}_{dd}$, $\widehat{S}_{dh}$, and $\widehat{S}_{hh}$ with Newey and West's (1987) HAC estimator, using a bandwidth of 2τ, where τ denotes

the forecast horizon.[8] This bandwidth setting allows for the possibility that noise in data revisions create some serial correlation in even one–step ahead forecast errors.[9] All test statistics are compared against critical values from the standard normal distribution. We report the percentage of 10,000 simulations in which the null of equal accuracy is rejected at the 5% significance level (using a critical value of ± 1.96).

Finally, with quarterly data in mind, we consider a range of sample sizes. For simplicity, we report results for a single R setting, of 80; results with $R = 40$ are very similar. We report results for four different P settings: 20, 40, 80, and 160.

4.2 Monte Carlo design: nested case

In the nested forecast case, we also consider three DGPs. For DGPs 1 and 2, the final data are generated by

$$\begin{aligned} y_t &= .7y_{t-1} + \beta_{22}x_{t-1} + e_{y,t} + v_{y,t} \qquad (16) \\ x_t &= .7x_{t-1} + e_{x,t} + v_{x,t} \\ Var\begin{pmatrix} e_{y,t} \\ e_{x,t} \\ v_{y,t} \\ v_{x,t} \end{pmatrix} &= \begin{pmatrix} .8 & & & \\ cov(e_y, e_x) & .2 & & \\ 0 & 0 & .2 & \\ 0 & 0 & 0 & .3 \end{pmatrix} \end{aligned}$$

In DGP 1 experiments with $\Omega > 0$, $cov(e_y, e_x) = .35$; in DGP 2, $cov(e_y, e_x) = .25$. In size experiments with noise but $\Omega = 0$, $cov(e_y, e_x) = 0$. The DGPs also differ in their parameterizations of the noise process, as described below. In size experiments, $\beta_{22} = 0$; in power experiments, $\beta_{22} = 0.3$.

For DGP 3, the final data are generated by

$$\begin{aligned} y_t &= -.4y_{t-1} - .3y_{t-2} - .2y_{t-3} + .1y_{t-4} + \beta_{22}x_{t-1} + e_{y,t} + v_{y,t} \qquad (17) \\ x_t &= 1.1x_{t-1} - .3x_{t-2} + e_{x,t} + v_{x,t} \\ Var\begin{pmatrix} e_{y,t} \\ e_{x,t} \\ v_{y,t} \\ v_{x,t} \end{pmatrix} &= \begin{pmatrix} .8 & & & \\ cov(e_y, e_x) & .2 & & \\ 0 & 0 & .2 & \\ 0 & 0 & 0 & .3 \end{pmatrix} \end{aligned}$$

In size experiments with noise but $\Omega = 0$, we set $cov(e_y, e_x)$ to 0. In all other experiments, $cov(e_y, e_x) = .25$. In size experiments, $\beta_{22} = 0$; in power experiments, $\beta_{22} = 0.3$.

[8] Using a bandwidth of $2(\tau - 1)$ yields very similar results.

[9] For example, if the true model is an AR(1) in y, with revisions generated by a process like that in (11), one–step ahead real–time forecast errors will contain an MA(1) component, due to the news innovation.

For all DGPs, the revision processes take the form:

$$\begin{aligned} y_t(t) &= y_t - v_{y,t} + w_{y,t} \\ x_t(t) &= x_t - v_{x,t} + w_{x,t} \\ & \quad w_{y,t}, w_{x,t} \; iid \; N(0, \cdot), \end{aligned} \tag{18}$$

Note that setting the variances of the w terms to zero yields revisions that are comprised entirely of news. In DGPs 1 and 2, the final estimates are released with a one–period delay; in DGP 3, the delay is four periods.

In DGP 1 experiments with noise, the noise innovation variances are set to $\sigma^2_{w,y} = 1.8$ and $\sigma^2_{w,x} = .5$. Under this parameterization, the correlation of the revision in y with the initial estimate is about -0.7, and the variance of revisions is about the same as the variance of the final data on y — well above what is observed in data on inflation. With $cov(e_y, e_x)$ = .35, this parameterization yields a relatively large Ω, taken as the baseline case. We also consider two variants. In the first, a news experiment, the settings are the same as in the baseline case, except that $\sigma^2_{w,y} = \sigma^2_{w,x} = 0$. In the second, an $\Omega = 0$ experiment, $\sigma^2_{w,y} =$ 1.8 and $\sigma^2_{w,x} = .5$ (as in the baseline), but $cov(e_y, e_x) = 0$.

The baseline experiments for DGPs 2 and 3 with noise use $\sigma^2_{w,y} = .2$ and $\sigma^2_{w,x} = .5$, which makes the correlation of the revision in y with the initial estimate about -0.25, and the variance of revisions in y about 20 to 30 percent of the variance of the final data on y — roughly in line with actual data. Analytically, we have verified that the DGP 2 parameterization using $cov(e_y, e_x) = .25$ yields a relatively small population Ω. We also consider two other versions of DGP 3, with the following variations on the baseline setting: a news experiment with $\sigma^2_{w,y} = \sigma^2_{w,x} = 0$ and an $\Omega = 0$ experiment with $cov(e_y, e_x) = 0$.

In DGP 1 and 2 experiments, we test for equal accuracy of τ-horizon forecasts from

$$y^{(\tau)}_{t+\tau} = a_1 y_t + u_{1,t+\tau} \tag{19}$$

$$y^{(\tau)}_{t+\tau} = a_2 y_t + b_2 x_t + u_{2,t+\tau}, \tag{20}$$

where $y^{(\tau)}_{t+\tau} \equiv \tau^{-1} \sum_{s=1}^{\tau} y_{t+s}$. In DGP 3 experiments, the forecasting models are

$$y^{(\tau)}_{t+\tau} = a_0 + a_1 y_t + a_2 y_{t-1} + a_3 y_{t-2} + a_4 y_{t-3} + u_{1,t+\tau} \tag{21}$$

$$y^{(\tau)}_{t+\tau} = b_0 + b_1 y_t + b_2 y_{t-1} + b_3 y_{t-2} + b_4 y_{t-3} + b_5 x_t + u_{2,t+\tau}. \tag{22}$$

At each forecast origin t, the observable time series for each variable consists of an initial or first vintage estimates for period $t - r + 1$ through t and final values for periods $t - r$

and earlier, where $r = 1$ in DGPs 1 and 2 and $r = 4$ in DGP 3. The parameters of the forecasting models are estimated recursively by OLS.

In evaluating forecasts, we compute forecast errors using actual values of y taken to be the initial estimate published in period t, $y_t(t)$. The null hypothesis is that the variables included in the larger model and not the smaller have no predictive content. To test this null, from the forecast errors we form the MSE-F test and various versions of the MSE-t test, and compare them to various sources of critical values. We reject the null if the test statistic exceeds the relevant right-tail critical value. We report the percentage of 10,000 simulations in which the null of equal accuracy is rejected at the 5% significance level.

More specifically, we construct the MSE-F test (8) and compare it against asymptotic critical values simulated as in Clark and McCracken (2005). We construct the conventional version of the MSE-t test, defined as MSE-$t(S_{dd}) = P^{1/2}(MSE_1 - MSE_2)/\widehat{S}_{dd}^{1/2}$, and compare it against both Clark and McCracken (2005) and standard normal critical values. Finally, we construct our proposed MSE-$t(\Omega)$ test, using the square root of $\widehat{\Omega} = 2\widehat{\Pi}\widehat{F}(-J\widehat{B}_1J' + \widehat{B}_2)\widehat{S}_{hh}(-J\widehat{B}_1J' + \widehat{B}_2)\widehat{F}'$ as the standard error, and compare it against standard normal critical values. We compute the long-run variances $\widehat{S}_{dd}$ and $\widehat{S}_{hh}$ with Newey and West's (1987) HAC estimator, using a bandwidth of 2τ, where τ denotes the forecast horizon.[10]

Finally, with quarterly data in mind, we consider a range of sample sizes. For simplicity, we report results for a single R setting, of 80; results with $R = 40$ are very similar. We report results for four different P settings: 20, 40, 80, and 160.

4.3 Monte Carlo results: non-nested case

Table 1 reports size results from non–nested forecast simulations. To establish a baseline, we first consider the properties of forecast tests in which revisions contain only news, in which case $F = 0$, and no standard error correction is needed.[11] We then consider results for noisy revisions, in which case $F \neq 0$ and, in principle, a standard error correction is necessary.

With revisions that contain only news, in results for DGPs 1-3 the size of the unadjusted t-test for equal MSE (MSE-$t(S_{dd})$) ranges from 6 to 28 percent — such that the test ranges from slightly to significantly oversized. With large forecast samples, the test tends to be

[10] Using a bandwidth of $2(\tau - 1)$ yields very similar results.

[11] Results for simulations with no data revisions are very similar to those reported for the news case.

close to correctly sized. But the size of the test rises as the sample shrinks and as the horizon increases.

Even though no standard error correction is asymptotically necessary, the adjusted t-test (MSE-$t(\Omega)$) seems to have better small-sample size properties, at least in smaller forecast samples. Across results for DGPs 1-3 with just news, the size of the adjusted test ranges from 4 to 14 percent — from slightly undersized to modestly oversized. These results indicate that, in a context in which a practitioner can't be sure that the revisions in his/her data set contain only news and not noise, applying our correction won't harm inference if the revisions contain only news, and may improve it.

Consider now the size of the tests in the case of predictable revisions (noise). In this case, the unadjusted MSE-t test might be expected to be oversized, more so for larger P/R than smaller P/R, because the variance in the test fails to account for (understate) the variance impact of the predictable revisions. In the case of one–step ahead forecasts from DGPs 1 and 2, we can analytically compute the population value of the correction $FBS_{dh} + FBS_{hh}BF'$ from (7), to be 2.34 for DGP 1 and -.28 for DGP 2, compared to population S_{dd} values of roughly 3.5. Accordingly, based on the one–step ahead asymptotics, we might expect the unadjusted and adjusted tests to both be about correctly sized in results for DGP 2, but the unadjusted to be oversized in results for DGP 1.

The noisy revisions results in Table 1 are consistent with these asymptotics. In DGP 1 results, the unadjusted test is modestly to significantly oversized, yielding a rejection rate between 11 and 12 percent at the one-step horizon and between 8 and 23 percent at the four-step horizon. The adjusted test has much better properties, ranging from slightly undersized to modestly oversized: the rejection rates range from 4 to 6 percent for one-step forecasts and 3 to 9 percent for four-step forecasts. In DGP 2 results, the size of the unadjusted test is consistently lower, sometimes significantly, while the size of the adjusted test is typically a bit higher than in the DGP 1 results. With DGP 2, the size of the unadjusted test ranges from 5 to 9 percent at the one-step horizon and from 5 to 21 percent at the four-step horizon. The size of the adjusted test falls between 5 and 8 percent for one-step forecasts and 3 and 11 percent for four-step forecasts. Results for DGP 3 are qualitatively similar to those for DGP 2; while we haven't analytically determined the population value of the standard error correction for DGP 3, simulated averages of the correction indicate it is quite small. Overall, the results for the noise revisions simulations are similar to those for the

news case in that, in small samples, the adjusted t-test generally has better small sample properties than the unadjusted test, even if, in population, the necessary correction is small.

Table 2 reports power results. Again, we first consider the properties of forecast tests in which revisions contain only news and then consider results for noisy revisions. With revisions that contain only news, the powers of the adjusted and unadjusted tests for one-step ahead forecasts are virtually identical. For example, with DGP 3, $P = 80$, and $\tau = 1$, both tests have power of 80 percent. At the four-step horizon, power is generally lower, and, for smaller samples (P), the power of the unadjusted test is often modestly greater than that of the adjusted test. For instance, with DGP 1, $P = 40$, and $\tau = 4$, the powers of the unadjusted and adjusted tests are 24 and 19 percent, respectively. On balance, though, there is little practical difference in the powers of the tests, except in quite small samples.

In results from experiments with predictable revisions (noise) in all variables, differences in power remain small or modest for small P, but overall power can be trivial. In DGPs 1 and 3, the power of the unadjusted test ranges from 8 to 24 percent; the power of the adjusted test ranges from 4 to 17 percent. Power is much better for DGP 2, with trivial differences across the unadjusted and adjusted tests, except with small P and the four-step horizon. With DGP 2, the powers of one–step ahead tests fall between 22 and 84 percent; the powers of the four-step tests range from 12 to 23 percent. To see why power can fall off so much with the introduction of noisy revisions, we can analytically determine the population difference in MSEs for DGPs 1 and 2. The population value of the one–step ahead, real–time forecast error for model i is $e_{y,t+1} + w_{y,t+1} + \beta_j x_{j,t} + \beta_i v_{x_i,t} - \beta_i w_{x_i,t}$, where β_i denotes the coefficient on $x_{i,t}$ in the DGP for y_{t+1}. It follows that the difference in population MSEs is

$$\text{MSE}_1 - \text{MSE}_2 = \left(\beta_2^2 - \beta_1^2\right)\left(\sigma_x^2 - \sigma_{w,x}^2 - \sigma_{v,x}^2\right). \tag{23}$$

As this makes clear, the news and noise components of the x variables reduce the difference in population MSE's. Accordingly, power is lower in experiments with news and noise than in experiments with just news. Power is lower in DGP 1 than in DGP 2 because noise is much larger in DGP 1 than in DGP 2.

4.4 Monte Carlo results: nested case

Table 3 reports size results from nested model forecast simulations. Consistent with our theoretical results for the impact of noisy revisions, in DGP 1 (for which we can analyti-

cally determine that Ω is large), the standard MSE-F and MSE-$t(S_{dd})$ statistics compared against Clark and McCracken (2005) critical values suffer large size distortions. The size of the MSE-F test ranges from 5 to 26 percent; the size of the MSE-$t(S_{dd})$ test ranges from 21 to 32 percent. Comparing MSE-$t(S_{dd})$ against standard normal critical values also yields significant oversizing, with size ranging from 10 to 24 percent. Comparing our proposed statistic MSE-$t(\Omega)$ against standard normal critical values yields much more accurate inference, with size between 8 and 10 percent at the one-step horizon and between 13 and 15 percent at the four-step horizon. Admittedly, at the four-step horizon, all of the tests are oversized; however, our proposed test fares at least slightly better than the others.

With DGPs 2 and 3, for which Ω is non-zero but small in the baseline noise parameterizations, it is less clear that any single test is more reliable than the others. In simulations of DGPs 2 and 3 with noise, the MSE-F test seems most reliable, with size ranging from 4 to 10 percent. The MSE-$t(S_{dd})$ test compared against Clark and McCracken (2005) critical values is less reliable (mostly so for small P or four-step forecasts), with size between 5 and 21 percent. Comparing the same test against standard normal critical values tends to yield an undersized test (except for small P and longer forecast horizons). Finally, our proposed MSE-$t(\Omega)$ test is consistently oversized, with size between 6 and 14 percent.

When revisions contain only news or contain noise but the DGP is such that $\Omega = 0$, the MSE-F test compared against Clark and McCracken (2005) critical values is closest to being correctly sized; our proposed MSE-$t(\Omega)$ is modestly oversized, in line with its performance in the baseline experiments. In simulations of DGP 3 with news, the size of MSE-F ranges from 5 to 9 percent; the size of MSE-$t(\Omega)$ varies from 8 to 15 percent.[12] In simulations of DGPs 1 and 3 with noise but $\Omega = 0$, the sizes of the MSE-F and MSE-$t(\Omega)$ tests range from, respectively, 2 to 9 percent and 9 to 15 percent. As in the baseline experiments, the MSE-$t(S_{dd})$ test compared against Clark and McCracken (2005) critical values tends to be more oversized than the MSE-F test (except at the largest P setting); the same test compared against standard normal critical values is generally undersized. Therefore, even though news-only revisions or noisy revisions with $\Omega = 0$ make the true asymptotic distributions more complicated than those developed in Clark and McCracken (2005) and this paper, it seems that, for some range of practical applications, the MSE-F and MSE-$t(\Omega)$ tests should yield reasonably reliable inference (under the null).

[12] Simulations of DGPs 1 and 2 with news yield similar results.

Table 4 provides power results from nested model forecast simulations. In the DGP 1 with noise experiment, the MSE-F and MSE-$t(\Omega)$ tests have comparable power: e.g., at the one-step horizon, the power of the former ranges from 39 to 98 percent; the power of the latter ranges from 63 to 93 percent (in relative terms, MSE-$t(\Omega)$ tends to be more powerful for smaller P and less powerful for larger P). Comparing the conventional MSE-$t(S_{dd})$ statistic against either Clark and McCracken (2005) or standard normal critical values typically yields lower power. For instance, at the one-step horizon, comparing MSE-$t(S_{dd})$ against Clark-McCracken critical values yields a rejection rate between 52 and 96 percent. In experiments for DGPs 2 and 3 with noise, power is generally much lower — often trivial for MSE-$t(S_{dd})$ compared against standard normal critical values. At the one-step horizon, the power of the MSE-F test ranges from 23 to 48 percent; the power of the MSE-$t(\Omega)$ test varies from 23 to 40 percent. In the same experiments, the power of MSE-$t(S_{dd})$ compared against standard normal critical values falls between 4 and 9 percent.

Perhaps not surprisingly, data revisions significantly impact the power of the tests. In simulations of DGP 3 with revisions containing only news, power is significantly higher than in the baseline case of noisy revisions. For example, the power of MSE-F for one-step ahead forecasts ranges from 42 to 88 percent in the news experiment, compared to a range of 26 to 34 percent in the noise experiment. Moreover, in simulations of DGP 2 and 3 using just the (same) final data to generate and evaluate forecasts (results not reported in Table 4), the tests show much better power. For example, in these revision–free DGP 2 experiments, the power of our proposed MSE-$t(\Omega)$ ranges from 52 to 84 percent; in revision-free DGP 3 simulations, the power of the same test varies from 68 to 96 percent. For both DGPs 2 and 3, data revisions significantly lower the average difference in the (real time) MSEs of models 1 and 2, to be roughly 0 (a result we have verified analytically).[13]

On balance, in the face of potentially predictable data revisions in nested model forecast comparisons, it would seem useful to consider results from multiple tests, preferably MSE-F and MSE-$t(\Omega)$. In cases in which Ω is large, our proposed test MSE-$t(\Omega)$ should be preferred, but in many practical settings, Ω seems likely to be small. In such settings, the MSE-F test compared against Clark and McCracken (2005) critical values — which is technically valid only in the absence of revisions — still seems to work reasonably despite

[13]However, data revisions do not ensure smaller differences in MSEs and lower power. The population difference in MSEs can be shown analytically to depend on not only the news and noise variances but also all other moments of the data. Under some DGP specifications, such as in the case of DGP 1, the difference in MSE can be greater with revisions than without.

the potential impact of revisions on the asymptotic distribution.

5 Application to Inflation Forecasting

In this section we use the tests and inference approaches described above to determine whether, in real time data, various measures of economic activity have predictive content for inflation. The inflation measure we forecast is the change in the inflation rate of the GDP price index. We consider one–quarter and one–year ahead forecasts of inflation from an AR model, a model including lags of the change in inflation and GDP growth, and a model including lags of the change in inflation and the output gap (HP detrended output). To illustrate non–nested testing, we compare forecasts from the model with GDP growth to the model with the output gap. To illustrate nested testing, we compare forecasts from the model with GDP growth to the AR model. Real-time evidence in Orphanides and van Norden (2005) suggests a model with GDP growth to be superior in inflation forecasting.

5.1 Data

Data on real output and the price index are taken from the Federal Reserve Bank of Philadelphia's Real–Time Data Set for Macroeconomists (RTDSM). For simplicity, we simply use the notation "GDP" and "GDP price index" to refer to the output and price series, even though the measures are based on GNP and a fixed weight deflator for much of the sample. The full forecast evaluation period runs from 1970:Q1 through 2003:Q4. For each forecast origin t in 1970:Q1 through 2003:Q4, we use the real time data vintage t to estimate output gaps, (recursively) estimate the forecast models, and then construct forecasts for periods t and beyond. The starting point of the model estimation sample is always $1961{:}1{+}\tau - 1$, where τ denotes the forecast horizon.

In evaluating real time forecast accuracy, we consider a range of possible definitions (vintages) of actual inflation. One estimate is the first one available in the RTDSM, one quarter after the end of the forecast observation date (i.e., inflation for period t published in period $t+1$). Another is the second estimate or vintage available in the RTDSM, published with a two–quarter delay.[14] We also consider estimates of inflation published with delays of five and 13 periods.

[14]Studies such as Romer and Romer (2000) use the second available estimates of the GDP/GNP deflator as actuals in evaluating forecast accuracy.

5.2 Models

Following Stock and Watson (1999, 2003) and Clark and McCracken (2006), among others, we obtain forecasts of the change in inflation at horizon τ from reduced–form Phillips curves:

$$\pi_{t+\tau}^{(\tau)} - \pi_t = \alpha_0 + \sum_{l=0}^{3} \alpha_l \Delta\pi_{t-l} + \beta x_t + u_{PC,t+\tau}, \tag{24}$$

where inflation is $\pi_t^{(\tau)} \equiv (400/\tau)\ln(p_t/p_{t-\tau})$, $\pi_t^{(1)} \equiv \pi_t$, and x_t is a measure of economic activity. In one version of this model, the x_t variable is defined as the four–quarter GDP growth rate, $100\ln(\text{GDP}_t/\text{GDP}_{t-4})$. In the other, x_t is defined as (100 times) HP-detrended log GDP.

In addition to comparing forecasts from one version of (24) with GDP growth to another with the output gap, we compare forecasts from the model with GDP growth to forecasts from the following AR model for the change in inflation:

$$\pi_{t+\tau}^{(\tau)} - \pi_t = \alpha_0 + \sum_{l=0}^{1} \alpha_l \Delta\pi_{t-l} + u_{AR,t+\tau}. \tag{25}$$

In computing the various versions of the MSE-t test, we use the Newey and West (1987) estimator of the long-run variances S_{dd}, S_{dh}, and S_{hh}, with a bandwidth of 2τ.[15]

5.3 Results

Table 5 presents results for the (non–nested) comparison of forecasts from the models with GDP growth (model 1) and the output gap (model 2). For most samples and definitions of actuals, although not all, the model with GDP growth yields slightly more accurate forecasts. The advantage of the model with GDP growth is consistently greater in year-ahead forecasts than one quarter-ahead forecasts. However, there is little evidence of statistical significance in the forecast accuracy differences. If the conventional variance $\widehat{S}_{dd}$ is used in forming the t-test, the null of equal accuracy is rejected only once at the one–step horizon (for the 1985-2003 sample using the inflation estimates published with a 13 period delay as actuals), but for all 1985-2003 samples at the one–year ahead horizon. Consistent with our Monte Carlo evidence, in most cases, taking account of the potential for predictability in the data revisions raises the estimated standard error. At the one–step horizon, though, the impact is pretty small in most cases, particularly in results for the 1970-2003 and 1970-84 samples. At the one–step horizon, the adjustment has a bigger impact in the 1985-2003

[15] Results are qualitatively the same with $2(\tau - 1)$.

results. Most notably, for the 1985-2003 sample using the inflation estimates published with a 13 period delay as actuals, the null of equal accuracy is not rejected based on the adjusted variance estimate, but it is when based on the unadjusted variance. The adjustment has a considerably bigger impact in the one–year ahead forecasts. The rejections of equal accuracy for all 1985-2003 samples based on the t-test using $\widehat{S}_{dd}$ go away when the test uses the adjusted variance $\widehat{\Omega}$.

Table 6 provides results for the (nested) comparison of forecasts from the AR(2) model (model 1) and the model with four lags of inflation and GDP growth (model 2). For nearly all samples and definitions of actuals, the forecasts from the model with GDP growth are more accurate than the AR(2) forecasts, slightly so at the one–step horizon and more substantially at the one–year horizon. When we abstract from the potential impact of predictable data revisions on test behavior, and compare MSE-F and MSE-$t(S_{dd})$ to asymptotic critical values simulated as in Clark and McCracken (2005), for most definitions of actual inflation we reject the null AR model with the full 1970-2003 sample of forecasts and the 1985-2003 sample. At the one-year horizon, the null is also always rejected for the 1970-84 sample. If the same MSE-$t(S_{dd})$ test is compared against standard normal critical values (1.282 for a one-sided 10% test), for one–step ahead forecasts the null is consistently rejected for the 1985-2003 sample but never rejected for the 1970-2003 period. For one–year ahead forecasts, the null AR model is nearly always rejected for the 1970-2003 and 1970-84 samples, but never for the 1985-2003 period. Taking account of data revisions by using the variance $\widehat{\Omega}$ in the MSE-t test increases the (absolute) value of the t–statistic in all but one case. However, in only two cases — one-step ahead forecasts for 1970-2003 and 1970-84 evaluated with first available estimates of inflation — is the adjusted t–statistic significant when the unadjusted t–statistic (compared against standard normal critical values) is not.

Overall, the two tests that the Monte Carlo evidence suggests to be most reliable in nested model comparisons — MSE-F and MSE-$t(\Omega)$ — are in general, although not universal, agreement. At the one-step horizon, both tests indicate GDP growth is useful for forecasting inflation from 1985 to 2003; the MSE-F test, but not MSE-$t(\Omega)$, indicate growth is helpful for forecasting in the 1970-2003 sample. At the four-step horizon, both tests consistently reject the null of equal accuracy for the 1970-2003 and 1970-1984 samples.

6 Conclusion

In this paper we derive the limiting distributions for tests of equal predictive ability when forecasting with real time vintage data. Specifically, we address the impact of revisions on the asymptotic distributions of the t–statistic for equal MSE between non-nested models developed by Diebold and Mariano (1995) and West (1996) and the F– and t–type tests of equal MSE between nested models developed in Clark and McCracken (2005) and McCracken (2006). We show that when revised data is used to construct and evaluate forecasts these tests typically do not have the same asymptotic distributions as when the data is never revised. With these new distributions in hand, we show how to conduct asymptotically valid inference. In the cases we consider, the tests are asymptotically standard normal and hence inference can be conducted using the relevant tables.

Using our asymptotics, we then conduct a range of Monte Carlo simulations to examine the finite–sample properties of the tests. Overall, these results broadly confirm our asymptotic approximations. In terms of size, ignoring the data revisions can produce oversized tests. Taking revisions into account by using our proposed tests can yield more reliable inferences, although in practice, there will be situations in which our proposed corrections are not very important. Data vintage also has an impact on the power of the tests. Typically, power is lower in data subject to revision than in data that are unrevised. The revisions drive a wedge between the properties of the dependent variable defining the predictive model and that used for evaluation. Depending on the exact relationships across vintages, predictive content for one vintage need not imply the same for another.

In the final part of our analysis, we illustrate the usage of our tests with an application to competing forecasts of U.S. inflation.

References

Aruoba, S.B. (2006), "Data Revisions Are Not Well-Behaved," *Journal of Money, Credit, and Banking*, forthcoming.

Chao, J., Corradi, V., Swanson, N. R. (2001), "An Out of Sample Test for Granger Causality," *Macroeconomic Dynamics*, 5, 598-620.

Clark, T.E., and McCracken, M.W. (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85-110.

Clark, T.E., and McCracken, M.W. (2005), "Evaluating Direct Multistep Forecasts," *Econometric Reviews*, 24, 369-404.

Clark, T.E., and McCracken, M.W. (2006), "The Predictive Content of the Output Gap for Inflation: Resolving In–Sample and Out–of–Sample Evidence," *Journal of Money, Credit, and Banking*, 38, 1127-1148.

Corradi, V., and Swanson, N.R. (2002), "A Consistent Test for Nonlinear Out–of–Sample Predictive Accuracy," *Journal of Econometrics*, 110, 353-381.

Corradi, V., Swanson, N.R., and Olivetti, C. (2001), "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 105, 315-358.

Croushore, D., and Stark, T. (2003), "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?" *The Review of Economics and Statistics*, 85, 605-617.

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253-263.

Faust, J., and Wright, J.H. (2005), "News and Noise in G-7 GDP Announcements," *Journal of Money, Credit, and Banking*, 37, 403-420.

Giacomini, R., and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545-1578.

Godfrey, L.G., and Pesaran, M.H. (1983), "Tests of Non–Nested Regression Models: Small Sample Adjustments and Monte Carlo Evidence," *Journal of Econometrics*, 21, 133-154.

Granger, C.W.J., and Newbold, P. (1977), *Forecasting Economic Time Series*, New York: Academic Press.

Howrey, E.P. (1978), "The Use of Preliminary Data in Econometric Forecasting," *Review of Economics and Statistics*, 60, 193-200.

Koenig, E.F., Dolmas, S., and Piger, J. (2003), "The Use and Abuse of Real-Time Data in Economic Forecasting," *The Review of Economics and Statistics*, 85, 618-628.

Mankiw, N.G., Runkle, D.E., and Shapiro, M.D. (1984), "Are Preliminary Announcements of the Money Stock Rational Forecasts?" *Journal of Monetary Economics*, 14, 15-27.

McCracken, M.W. (2000), "Robust Out–of–Sample Inference," *Journal of Econometrics*, 99, 195-223.

McCracken, M.W. (2006), "Asymptotics for Out–of–Sample Tests of Causality," *Journal of Econometrics*, forthcoming.

Newey, W.K., and West, K.D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703-708.

Orphanides, A., and van Norden, S. (2005), "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time," *Journal of Money, Credit, and Banking*, 37, 583-601.

Robertson, J.C., and Tallman, E.W. (1998), "Data Vintages and Measuring Forecast Model Performance," *Federal Reserve Bank of Atlanta Economic Review*, 83 (Fourth Quarter), 4-20.

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

Rossi, B. (2005), "Testing Long-Horizon Predictive Ability with High Persistence, and the Meese-Rogoff Puzzle," *International Economic Review*, 46, 61-92.

Stock, J.H., and Watson, M.W. (1999), "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293-335.

Stock, J.H., and Watson, M.W. (2003), "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature*, 41, 788-829.

Swanson, N.R. (1996), "Forecasting Using First Available Versus Fully Revised Economic Time Series Data," *Studies in Nonlinear Dynamics and Econometrics*, 1, 47-64.

Vuong, Q.H. (1989), "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307-333.

West, K.D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067-1084.

West, K.D., and McCracken, M.W. (1998), "Regression–Based Tests of Predictive Ability," *International Economic Review*, 39, 817-840.

7 Appendix 1: Theory Details

Most results follow from very similar arguments to those in West (1996), McCracken (2000), and Clark and McCracken (2005) but keeping track of the fact that while $(x'_{i,t}, y_{t+\tau})'$ is covariance stationary, it need not have the same first and second moments as $(x'_{i,t}(t), y_{t+\tau}(t'))'$ due to the revision process.

In order to keep track of this distinction, some notation is useful: $B_i(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}x'_{i,s})^{-1}$, $\hat{B}_i(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}(t)x'_{i,s}(t))^{-1}$, $G_i(t) = t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}y_{s+\tau}$, $\hat{G}_i(t) = t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}(t)y_{s+\tau}(t)$, $H_i(t) = t^{-1}\sum_{s=1}^{t-\tau}(y_{s+\tau} - x'_{i,s}\beta_i^*)x_{i,s}$, and $\hat{H}_i(t) = t^{-1}\sum_{s=1}^{t-\tau}(y_{s+\tau}(t) - x'_{i,s}(t)\beta_i^*)x_{i,s}(t)$. If we let $\hat{H}_i(t) - H_i(t) = t^{-1}v_{i,t}$ we obtain the identity $\hat{\beta}_{i,t} \equiv \hat{B}_i(t)\hat{G}_i(t) = \beta_i^* + B_i(t)H_i(t) + B_i(t)(t^{-1}v_{i,t}) + (\hat{B}_i(t) - B_i(t))G_i(t) + (\hat{B}_i(t) - B_i(t))(t^{-1}v_{i,t})$. In addition, we let $\sup_t$ denote $\sup_{R\le t\le T}$, and for any matrix A with elements $a_{i,j}$, $|A|$ denotes $\max_{i,j}|a_{i,j}|$. Finally, ignoring the finite sample distinction between summing over P and $P-\tau+1$ elements, each set of results are based upon the decomposition of $\bar{d}$ into four bracketed $\{.\}$ terms.

$$
\begin{aligned}
\bar{d} &= \{P^{-1}\textstyle\sum_{t=R}^{T}(u^2_{1,t+\tau}(t') - u^2_{2,t+\tau}(t'))\} \\
&+ \{P^{-1}\textstyle\sum_{t=R}^{T}(2\sum_{i=1}^{2}(-1)^i h_{i,t+\tau}(t')B_i(t)H_i(t))\} \\
&+ \{P^{-1}\textstyle\sum_{t=R}^{T}(\sum_{i=1}^{2}(-1)^{i+1}H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)H_i(t))\} \\
&+ \{P^{-1}\textstyle\sum_{t=R}^{T}(\sum_{i=1}^{2}(-1)^i(2h_{i,t+\tau}(t')B_i(t)(t^{-1}v_{i,t}) + 2h_{i,t+\tau}(t')(\hat{B}_i(t) - B_i(t))\hat{H}_i(t) \\
&- H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)(t^{-1}v_{i,t}) - 2H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)(\hat{B}_i(t) - B_i(t))\hat{H}_i(t) \\
&- (t^{-1}v'_{i,t})B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)(t^{-1}v_{i,t}) - 2(t^{-1}v'_{i,t})B_i(t)x_{i,t}(t)x'_{i,t}(t)(\hat{B}_i(t) - B_i(t))\hat{H}_i(t) \\
&- \hat{H}'_i(t)(\hat{B}_i(t) - B_i(t))x_{i,t}(t)x'_{i,t}(t)(\hat{B}_i(t) - B_i(t))\hat{H}_i(t)))\}
\end{aligned} \tag{26}
$$

Proof of Lemma 1: Given the decomposition in (26), it suffices to show that the $P^{1/2}$-scaled second bracketed term equals $FB(P^{-1/2}\sum_{t=R}^{T}H(t)) + o_p(1)$ and the $P^{1/2}$-scaled third and fourth bracketed terms are $o_p(1)$. That the first term equals $FB(P^{-1/2}\sum_{t=R}^{T}H(t)) + o_p(1)$ follows from algebra nearly identical to that in West (1996, Lemma 4.1; see apx. Lemma A4). Proofs that the third term, and each component of the fourth term are $o_p(1)$ follow similar logic. For example,

$$
\begin{aligned}
&|P^{-1/2}\textstyle\sum_{t=R}^{T}H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)H_i(t)| \le \\
&k^8(P^{-1/2})(P^{-1}\textstyle\sum_{t=R}^{T}|x_{i,t}(t)x'_{i,t}(t)|)(\sup_t|B_i(t)|)^2(\sup_t|P^{1/2}H_i(t)|)^2
\end{aligned}
$$

and

$$
\begin{aligned}
&|P^{-1/2}\textstyle\sum_{t=R}^{T}h'_{i,t+\tau}(t')B_i(t)(t^{-1}v_{i,t})| \le \\
&k^2(P^{1/2}/R)(\sup_t|B_i(t)|)(P^{-1}\textstyle\sum_{t=R}^{T}|h_{i,t+\tau}(t')||v_{i,t}|)
\end{aligned}.
$$

Since Assumption 2 suffices for $P^{-1}\sum_{t=R}^{T}|x_{i,t}(t)x'_{i,t}(t)|$, $\sup_t|B_i(t)|$, $\sup_t|P^{1/2}H_i(t)|$, and $P^{-1}\sum_{t=R}^{T}|h_{i,t+\tau}(t')||v_{i,t}|$ to each be $O_p(1)$, Assumption 4 or 4′ imply each term is $o_p(1)$.

Proof of Theorem 1: Given Lemma 1 the result follows nearly identical logic to that in West (1996) Theorem 4.1.

Proof of Lemma 2: First note that, under the null, the initial bracketed term in (26) is zero since $u_{1,t+\tau}(t') = u_{2,t+\tau}(t') = u_{t+\tau}(t')$. (i) If we also note that $J'x_{2,t}(t) = x_{1,t}(t)$ so that $J'H_2(t) = H_1(t)$, and take account of the (nested model) definition of F, the proof is identical to that in Lemma 1.

(ii) The key differences in the proof are (a) the scaling by $R^{1/2}$ rather than $P^{1/2}$, and (b) given our assumptions, $\sup_t R^{1/2}|B_i(t)H_i(t) - B_i(R)H_i(R)| = o_p(1)$; a result that follows from Lemma 8 of Clark and McCracken (2001). With this tool in hand we first show that the second term in (26) equals $F(-JB_1J' + B_2)(R^{1/2}H(R)) + o_p(1)$. To do so note that

$$\begin{array}{l} R^{1/2}P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t')B_i(t)H_i(t) = R^{1/2}(P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t'))B_i(R)H_i(R) \\ +R^{1/2}(P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t')(B_i(t)H_i(t) - B_i(R)H_i(R))) \end{array} .$$

Since $B_i(R) \to_p B_i$, $P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t') \to_p Eh'_{i,t+\tau}(t')$, $R^{1/2}H_i(R) = O_p(1)$, and

$$\begin{array}{l} |R^{1/2}(P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t')(B_i(t)H_i(t) - B_i(R)H_i(R)))| \leq \\ k(P^{-1}\sum_{t=R}^{T} |h'_{i,t+\tau}(t')|)(\sup_t R^{1/2}|B_i(t)H_i(t) - B_i(R)H_i(R)|) \end{array}$$

we obtain the desired result.

Proofs that the third term, and each component of the fourth term are $o_p(1)$ follow logic comparable to that in Lemma 1 but adjusting for the rescaling. Using the same examples from Lemma 1,

$$\begin{array}{l} |R^{1/2}P^{-1}\sum_{t=R}^{T} H'_i(t)B_i(t)x_{i,t}(t)x'_{i,t}(t)B_i(t)H_i(t)| \leq \\ k^4(R^{-1/2})(P^{-1}\sum_{t=R}^{T} |x_{i,t}(t)x'_{i,t}(t)|)(\sup_t |B_i(t)|)^2(\sup_t |R^{1/2}H_i(t)|)^2 \end{array}$$

and

$$\begin{array}{l} |R^{1/2}P^{-1}\sum_{t=R}^{T} h'_{i,t+\tau}(t')B_i(t)(t^{-1}v_{i,t})| \leq \\ k(R^{-1/2})(\sup_t |B_i(t)|)(P^{-1}\sum_{t=R}^{T} |h'_{i,t+\tau}(t')||v_{i,t}|) \end{array} .$$

Since Assumption 2 suffices for $P^{-1}\sum_{t=R}^{T} |x_{i,t}(t)x'_{i,t}(t)|$, $\sup_t |B_i(t)|$, $\sup_t |R^{1/2}H_i(t)|$, and $P^{-1}\sum_{t=R}^{T} |h'_{i,t+\tau}(t')||v_{i,t}|$ to each be $O_p(1)$, Assumption 4′ implies each term is $o_p(1)$.

Proof of Theorem 2: (i) Given Lemma 2 (i) the result follows nearly identical logic to that in West (1996) Theorem 4.1. (ii) Given Lemma 2 (ii), and the fact that $R^{1/2}H(R) \to_d N(0, S_{hh})$, the result is immediate.

Proof of Theorem 3: (i) Assumption 2 suffices for $\hat{B}_i \to_p B_i$. That $\hat{F} \to_p F$ follows nearly identical arguments to that in Lemma 5.1 of West (1996). That $\hat{\Gamma}_{hh} \to_p \Gamma_{hh}$ is immediate from Theorem 5.1 of West (1996). Consider $\hat{\Gamma}_{dd}(j)$; $\hat{\Gamma}_{dh}(j)$ can be handled similarly. By adding and subtracting terms we have $\hat{\Gamma}_{dd}(j) = P^{-1}\sum_{t=R+j}^{T} d_{t+\tau}(t')d_{t+\tau-j}(t'-j) + r_T$ where

$$\begin{array}{l} r_T = -\bar{d}^2 + P^{-1}\sum_{t=R+j}^{T} (\hat{d}_{t+\tau}(t') - d_{t+\tau}(t'))d_{t+\tau-j}(t'-j) \\ +P^{-1}\sum_{t=R+j}^{T} d_{t+\tau}(t')(\hat{d}_{t+\tau-j}(t'-j) - d_{t+\tau-j}(t'-j)) \\ +P^{-1}\sum_{t=R+j}^{T} (\hat{d}_{t+\tau}(t') - d_{t+\tau}(t'))(\hat{d}_{t+\tau-j}(t'-j) - d_{t+\tau-j}(t'-j)) \end{array} .$$

Since Assumption 2 suffices for the first term to converge in probability to $\Gamma_{dd}(j)$, it is sufficient to show that for $m \in (0, .5)$, $r_T = o_p(P^{-m})$. That $P^m \bar{d}^2 = o_p(1)$ is immediate from either Theorems 1 or 2. The proof for each of the remaining components is similar. As an example note that the second component of r_T satisfies

$$\begin{array}{l} P^m|P^{-1}\sum_{t=R+j}^{T}(\hat{d}_{t+\tau}(t') - d_{t+\tau}(t'))d_{t+\tau-j}(t'-j)| \\ \leq 4k^2 \max_{i=1,2}[(\sup_t P^m|\hat{\beta}_{i,t} - \beta_i^*|)(P^{-1}\sum_{t=R}^{T}|h'_{i,t+\tau}(t')||d_{t+\tau-j}(t'-j)|)] \\ +2k^4 P^{-m} \max_{i=1,2}[(\sup_t P^m|\hat{\beta}_{i,t} - \beta_i^*|)^2(P^{-1}\sum_{t=R}^{T}|x_{i,t}(t)x'_{i,t}(t)||d_{t+\tau-j}(t'-j)|)] \end{array} .$$

Assumption 2 suffices for both $P^{-1}\sum_{t=R}^{T}|h'_{i,t+\tau}(t')||d_{t+\tau-j}(t'-j)|$ and $P^{-1}\sum_{t=R}^{T}|x_{i,t}(t)x'_{i,t}(t)||d_{t+\tau-j}(t'-j)|$ to be $O_p(1)$. Since $(\sup_t P^m|\hat{\beta}_{i,t} - \beta_i^*|) = o_p(1)$ follows from nearly identical arguments to that in Lemma A.3 of McCracken (2000), we obtain the desired result.

(ii) Given part (i)–and especially that $r_T = o_p(P^{-m})$–the proof is identical to that for Theorem 2.3.2 in McCracken (2000).

Table 1. Non-Nested Model Size Results

($R = 80$, nominal size = 5%)

test	**P = 20**	**40**	**80**	**160**	**P = 20**	**40**	**80**	**160**
	horizon = 1				**horizon = 4**			
	DGP 1, news							
MSE-$t(S_{dd})$	.10	.07	.06	.06	.22	.12	.08	.06
MSE-$t(\Omega)$	.08	.06	.05	.05	.11	.06	.04	.04
	DGP 2, news							
MSE-$t(S_{dd})$	.10	.07	.06	.06	.22	.12	.08	.06
MSE-$t(\Omega)$	.08	.06	.05	.05	.11	.06	.04	.04
	DGP 3, news							
MSE-$t(S_{dd})$	.10	.08	.07	.06	.28	.16	.11	.08
MSE-$t(\Omega)$	.05	.04	.04	.05	.14	.09	.06	.06
	DGP 1, noise							
MSE-$t(S_{dd})$	.11	.11	.11	.12	.23	.14	.10	.08
MSE-$t(\Omega)$	.06	.04	.04	.04	.09	.05	.03	.03
	DGP 2, noise							
MSE-$t(S_{dd})$	.09	.07	.06	.05	.21	.12	.07	.05
MSE-$t(\Omega)$	.08	.06	.05	.05	.11	.06	.04	.03
	DGP 3, noise							
MSE-$t(S_{dd})$	.10	.07	.06	.06	.27	.15	.09	.07
MSE-$t(\Omega)$	.05	.04	.04	.04	.14	.08	.05	.05

Notes:

1. DGPs 1 and 2 are given in equations (9) and (11). DGP 3 is given in equations (10) and (11). In these size experiments, the DGP coefficient β is set to 0. The DGP 1-2 forecasting models are given in equations (12) and (13); the DGP 3 models are given in equations (14) and (15).
2. R defines the size of the sample used to generate the first forecast. P defines the number of observations in the forecast sample. The number of Monte Carlo replications is 10,000.
3. MSE-$t(S_{dd})$ refers to an unadjusted t-test for equal MSE, using the conventional variance $\hat{S}_{dd}$. MSE-$t(\Omega)$ refers to an adjusted t-test for equal MSE, using the variance $\hat{\Omega} = \hat{S}_{dd} + 2\hat{\Pi}(\hat{F}\hat{B}\hat{S}_{dh} + \hat{F}\hat{B}\hat{S}_{hh}\hat{B}\hat{F}')$. All test statistics are compared against standard normal critical values of ± 1.96.

Table 2. Non-Nested Model Power Results

($R = 80$, nominal size = 5%)

test	**P = 20**	**40**	**80**	**160**	**P = 20**	**40**	**80**	**160**
	horizon = 1				**horizon = 4**			
	DGP 1, news							
MSE-$t(S_{dd})$	.70	.94	1.00	1.00	.26	.24	.32	.53
MSE-$t(\Omega)$	.70	.94	1.00	1.00	.20	.19	.28	.52
	DGP 2, news							
MSE-$t(S_{dd})$	.48	.75	.95	1.00	.26	.23	.30	.50
MSE-$t(\Omega)$	.48	.74	.96	1.00	.20	.18	.26	.48
	DGP 3, news							
MSE-$t(S_{dd})$	.34	.52	.80	.97	.33	.29	.38	.60
MSE-$t(\Omega)$	.32	.52	.80	.97	.25	.27	.38	.62
	DGP 1, noise							
MSE-$t(S_{dd})$	.12	.09	.10	.11	.22	.14	.10	.08
MSE-$t(\Omega)$	.09	.06	.06	.07	.13	.07	.05	.04
	DGP 2, noise							
MSE-$t(S_{dd})$	.22	.33	.54	.83	.23	.16	.16	.22
MSE-$t(\Omega)$	.22	.33	.55	.84	.17	.12	.14	.22
	DGP 3, noise							
MSE-$t(S_{dd})$	.11	.10	.11	.14	.24	.15	.11	.10
MSE-$t(\Omega)$	.10	.09	.11	.14	.17	.12	.10	.11

Notes:

1. In these power experiments, the DGP coefficient β is set to 0.6 in DGP 1-2 experiments and 0.75 in DGP 3 experiments.
2. See the notes to Table 1.

Table 3. Nested Model Size Results

($R = 80$, nominal size = 5%)

test	c.v.	P = 20	40	80	160	P = 20	40	80	160
		horizon = 1				horizon = 4			
		DGP 1, noise							
MSE-F	CM	.05	.11	.18	.26	.06	.11	.17	.25
MSE-$t(S_{dd})$	CM	.21	.23	.27	.32	.25	.22	.23	.26
MSE-$t(S_{dd})$	N	.10	.12	.17	.24	.16	.13	.14	.17
MSE-$t(\Omega)$	N	.10	.09	.09	.08	.15	.14	.13	.13
		DGP 2, noise							
MSE-F	CM	.04	.05	.06	.07	.04	.05	.06	.05
MSE-$t(S_{dd})$	CM	.12	.08	.06	.05	.19	.11	.06	.03
MSE-$t(S_{dd})$	N	.04	.02	.02	.01	.10	.05	.02	.01
MSE-$t(\Omega)$	N	.10	.10	.08	.07	.12	.11	.08	.06
		DGP 3, noise							
MSE-F	CM	.05	.07	.09	.10	.05	.07	.08	.09
MSE-$t(S_{dd})$	CM	.13	.10	.08	.07	.21	.14	.10	.06
MSE-$t(S_{dd})$	N	.04	.03	.02	.02	.12	.06	.03	.02
MSE-$t(\Omega)$	N	.12	.11	.10	.09	.14	.14	.11	.09
		DGP 3, news							
MSE-F	CM	.05	.06	.07	.07	.06	.08	.09	.08
MSE-$t(S_{dd})$	CM	.12	.09	.06	.04	.22	.14	.09	.05
MSE-$t(S_{dd})$	N	.04	.03	.02	.01	.13	.06	.03	.01
MSE-$t(\Omega)$	N	.11	.10	.09	.08	.15	.14	.11	.09
		DGP 1, noise, $\Omega = 0$							
MSE-F	CM	.02	.03	.05	.07	.03	.06	.07	.09
MSE-$t(S_{dd})$	CM	.13	.10	.09	.08	.22	.15	.11	.08
MSE-$t(S_{dd})$	N	.05	.04	.03	.03	.13	.08	.04	.03
MSE-$t(\Omega)$	N	.10	.10	.10	.10	.14	.13	.11	.10
		DGP 3, noise, $\Omega = 0$							
MSE-F	CM	.05	.07	.08	.08	.05	.07	.08	.08
MSE-$t(S_{dd})$	CM	.13	.10	.08	.05	.21	.14	.09	.06
MSE-$t(S_{dd})$	N	.04	.03	.02	.01	.12	.06	.03	.02
MSE-$t(\Omega)$	N	.11	.11	.10	.09	.15	.13	.11	.10

Notes:

1. DGPs 1 and 2 are given in equations (16) and (18). DGP 3 is given in equations (17) and (18). In these size experiments, the DGP coefficient β_{22} is set to 0. In the experiments for DGP 1 and DGP 3 with $\Omega = 0$, the error covariance $cov(e_y, e_x)$ is set to 0. The DGP 1-2 forecasting models are given in equations (19) and (20); the DGP 3 models are given in equations (21) and (22).

2. R defines the size of the sample used to generate the first forecast. P defines the number of observations in the forecast sample. The number of Monte Carlo replications is 10,000.

3. MSE-F refers to an F-test for equal MSE (given in equation 8). MSE-$t(S_{dd})$ refers to a t-test for equal MSE (given in equation (1)), using the conventional variance $\hat{S}_{dd}$. MSE-$t(\Omega)$ refers to a t-test for equal MSE using the variance $\hat{\Omega} = 2\hat{\Pi}\hat{F}(-J\hat{B}_1J' + \hat{B}_2)\hat{S}_{hh}(-J\hat{B}_1J' + \hat{B}_2)\hat{F}'$. The second column indicates what critical value is used. A 'CM' means the critical value is obtained from the simulation method of Clark and McCracken (2005); an 'N' means the critical value is taken from the standard normal distribution (1.645). All tests are one-sided, with the null rejected if the statistic exceeds the right-tail critical value.

Table 4. Nested Model Power Results

($R = 80$, nominal size = 5%)

test	**c.v.**	**P = 20**	**40**	**80**	**160**	**P = 20**	**40**	**80**	**160**
		horizon = 1				**horizon = 4**			
		DGP 1, noise							
MSE-F	CM	.39	.64	.87	.98	.25	.45	.70	.91
MSE-$t(S_{dd})$	CM	.52	.66	.83	.96	.48	.53	.65	.83
MSE-$t(S_{dd})$	N	.28	.41	.63	.88	.33	.34	.44	.65
MSE-$t(\Omega)$	N	.63	.72	.83	.93	.50	.56	.67	.81
		DGP 2, noise							
MSE-F	CM	.23	.31	.39	.48	.10	.16	.22	.28
MSE-$t(S_{dd})$	CM	.20	.20	.21	.23	.24	.19	.16	.15
MSE-$t(S_{dd})$	N	.08	.07	.07	.09	.14	.08	.06	.05
MSE-$t(\Omega)$	N	.38	.38	.39	.40	.31	.29	.28	.26
		DGP 3, noise							
MSE-F	CM	.26	.29	.33	.34	.15	.21	.29	.36
MSE-$t(S_{dd})$	CM	.18	.16	.14	.13	.27	.22	.21	.20
MSE-$t(S_{dd})$	N	.07	.06	.05	.04	.16	.11	.08	.08
MSE-$t(\Omega)$	N	.36	.32	.28	.23	.34	.34	.33	.32
		DGP 3, news							
MSE-F	CM	.42	.58	.75	.88	.19	.32	.48	.66
MSE-$t(S_{dd})$	CM	.35	.40	.51	.68	.36	.34	.37	.47
MSE-$t(S_{dd})$	N	.17	.19	.27	.42	.23	.18	.18	.24
MSE-$t(\Omega)$	N	.58	.64	.73	.83	.44	.48	.55	.65

Notes:

1. In these power experiments, the DGP coefficient β_{22} is set to 0.3.
2. See the notes to Table 3.

Table 5. Results for Non-Nested Model Inflation Forecasts

sample	MSE_1	MSE_2	MSE_1 - MSE_2	$\sqrt{S_{dd}/P}$	$\sqrt{\Omega/P}$	MSE-$t(S_{dd})$	MSE-$t(\Omega)$
1-step horizon							
actual inflation$_t$ = estimate published in $t+1$							
1970-2003	2.164	2.181	-.017	.173	.189	-.099	-.090
1970-1984	3.791	3.758	.033	.389	.396	.084	.083
1985-2003	.880	.937	-.056	.044	.070	-1.276	-.803
actual inflation$_t$ = estimate published in $t+2$							
1970-2003	2.311	2.372	-.061	.174	.181	-.353	-.339
1970-1984	4.033	4.073	-.040	.390	.378	-.104	-.107
1985-2003	.951	1.029	-.078	.051	.072	-1.523	-1.084
actual inflation$_t$ = estimate published in $t+5$							
1970-2003	2.481	2.447	.034	.211	.203	.162	.168
1970-1984	4.489	4.314	.174	.470	.416	.370	.419
1985-2003	.896	.972	-.076	.048	.072	-1.609	-1.055
actual inflation$_t$ = estimate published in $t+13$							
1970-2003	2.252	2.438	-.186	.157	.202	-1.189	-.922
1970-1984	4.196	4.512	-.315	.350	.431	-.902	-.731
1985-2003	.717	.801	-.084	.036	.069	*-2.364*	-1.215
4-step horizon							
actual inflation$_t$ = estimate published in $t+1$							
1970-2003	1.640	1.931	-.291	.266	.379	-1.092	-.767
1970-1984	2.939	3.257	-.317	.594	.797	-.534	-.398
1985-2003	.665	.937	-.271	.144	.227	*-1.877*	-1.192
actual inflation$_t$ = estimate published in $t+2$							
1970-2003	2.022	2.268	-.246	.269	.364	-.914	-.675
1970-1984	3.844	4.060	-.216	.598	.781	-.361	-.277
1985-2003	.655	.924	-.268	.152	.232	*-1.765*	-1.159
actual inflation$_t$ = estimate published in $t+5$							
1970-2003	1.937	2.262	-.325	.272	.393	-1.196	-.828
1970-1984	3.715	4.144	-.429	.605	.840	-.709	-.511
1985-2003	.604	.851	-.247	.143	.222	*-1.728*	-1.114
actual inflation$_t$ = estimate published in $t+13$							
1970-2003	1.997	2.468	-.471	.280	.451	*-1.683*	-1.044
1970-1984	3.841	4.599	-.758	.608	.966	-1.246	-.785
1985-2003	.614	.870	-.256	.127	.228	*-2.019*	-1.121

Notes:
1. The table compares the accuracy of real-time forecasts of the change in GDP inflation, from equation (24). Model 1 uses x_t = four-quarter GDP growth; Model 2 uses x_t = the output gap, computed with the HP filter. The models are estimated recursively, with the sample beginning in 1961:1+τ-1.
2. The MSEs are defined as annualized percentage points. MSE_1 refers to the mean square error of forecasts from the model with GDP growth; MSE_2 refers to the mean square error of forecasts from the model with the output gap. The MSEs are based on forecasts computed with various definitions of actual inflation used in computing forecast errors. The first panel takes actual to be the first available estimate of inflation; the next the second available estimate; and so on.
3. The variance Ω is defined as $\hat{S}_{dd} + 2\hat{\Pi}(\hat{F}\hat{B}\hat{S}_{dh} + \hat{F}\hat{B}\hat{S}_{hh}\hat{B}\hat{F}')$. The columns MSE-$t(S_{dd})$ and MSE-$t(\Omega)$ report t-statistics for the difference in MSEs computed with the variances $\hat{S}_{dd}$ and $\hat{\Omega}$, respectively. Test statistics rejecting the null of equal accuracy at a significance level of 10% or better are reported in a *slanted* font.

Table 6. Results for Nested Model Inflation Forecasts

sample	MSE_1	MSE_2	MSE_1 - MSE_2	$\sqrt{S_{dd}/P}$	$\sqrt{\Omega/P}$	MSE-$t(S_{dd})$	MSE-$t(\Omega)$	MSE-F
1-step horizon								
actual inflation$_t$ = estimate published in $t+1$								
1970-2003	2.368	2.164	.203	.200	.096	*1.019*	*2.118*	*12.786*
1970-1984	4.096	3.791	.305	.446	.214	*.684*	*1.426*	*4.829*
1985-2003	1.003	.880	.123	.057	.059	*2.162*	*2.093*	*10.644*
actual inflation$_t$ = estimate published in $t+2$								
1970-2003	2.359	2.311	.048	.156	.087	*.310*	.553	*2.841*
1970-1984	3.986	4.033	-.047	.347	.275	-.134	-.169	-.692
1985-2003	1.074	.951	.123	.047	.069	*2.608*	*1.777*	*9.834*
actual inflation$_t$ = estimate published in $t+5$								
1970-2003	2.565	2.481	.085	.188	.101	*.452*	.835	*4.646*
1970-1984	4.504	4.489	.016	.420	.281	.038	.056	.211
1985-2003	1.035	.896	.139	.047	.045	*2.947*	*3.070*	*11.813*
actual inflation$_t$ = estimate published in $t+13$								
1970-2003	2.297	2.252	.045	.176	.093	*.255*	.482	*2.713*
1970-1984	4.221	4.196	.025	.397	.261	.062	.094	.351
1985-2003	.778	.717	.061	.036	.013	*1.701*	*4.539*	*6.470*
4-step horizon								
actual inflation$_t$ = estimate published in $t+1$								
1970-2003	2.227	1.640	.587	.352	.117	*1.665*	*5.028*	*47.578*
1970-1984	4.271	2.939	1.332	.692	.416	*1.925*	*3.200*	*25.830*
1985-2003	.693	.665	.028	.130	.037	.213	.748	3.161
actual inflation$_t$ = estimate published in $t+2$								
1970-2003	2.676	2.022	.654	.423	.140	*1.545*	*4.689*	*43.038*
1970-1984	5.340	3.844	1.497	.855	.490	*1.749*	*3.055*	*22.194*
1985-2003	.678	.655	.022	.131	.042	.171	.533	2.605
actual inflation$_t$ = estimate published in $t+5$								
1970-2003	2.574	1.937	.637	.414	.131	*1.538*	*4.867*	*43.730*
1970-1984	5.148	3.715	1.434	.839	.430	*1.710*	*3.332*	*22.001*
1985-2003	.644	.604	.039	.143	.024	.275	*1.665*	*4.954*
actual inflation$_t$ = estimate published in $t+13$								
1970-2003	2.480	1.997	.483	.386	.110	*1.251*	*4.398*	*32.169*
1970-1984	5.046	3.841	1.205	.788	.423	*1.528*	*2.847*	*17.879*
1985-2003	.556	.614	-.058	.131	.054	-.444	-1.083	-7.200

Notes:
1. The table compares the accuracy of real-time forecasts of the change in GDP inflation, from equations (24) (MSE_1) and (25) (MSE_2), with x_t measured as four-quarter GDP growth. The models are estimated recursively, with the sample beginning in 1961:1+τ-1.
2. The MSEs are based on forecasts computed with various definitions of actual inflation used in computing forecast errors. The first panel takes actual to be the first available estimate of inflation; the next the second available estimate; and so on.
3. The variance Ω is defined as $2\hat{\Pi}\hat{I}\hat{F}(-J\hat{B}_1J' + \hat{B}_2)\hat{S}_{hh}(-J\hat{B}_1J' + \hat{B}_2)\hat{F}'$. The columns MSE-$t(S_{dd})$ and MSE-$t(\Omega)$ report t-statistics for the difference in MSEs computed with the variances $\hat{S}_{dd}$ and $\hat{\Omega}$, respectively. Test statistics rejecting the null of equal accuracy at a significance level of 10% or better are reported in a *slanted* font, for: MSE-$t(S_{dd})$ and MSE-F vs. critical values simulated as in Clark and McCracken (2005); and MSE-$t(\Omega)$ vs. standard normal critical values.