RESEARCH DIVISION

Working Paper Series

Testing for Unconditional Predictive Ability

Todd E. Clark
and
Michael W. McCracken

Working Paper 2010-031A
https://doi.org/10.20955/wp.2010.031

September 2010

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
P.O. Box 442
St. Louis, MO 63166

Testing for Unconditional Predictive Ability *

Todd E. Clark
Federal Reserve Bank of Kansas City

Michael W. McCracken
Federal Reserve Bank of St. Louis

April 2010

Abstract

This chapter provides an overview of pseudo-out-of-sample tests of unconditional predictive ability. We begin by providing an overview of the literature, including both empirical applications and theoretical contributions. We then delineate two distinct methodologies for conducting inference: one based on the analytics in West (1996) and the other based on those in Giacomini and White (2006). These two approaches are then carefully described in the context of pairwise tests of equal forecast accuracy between two models. We consider both non-nested and nested comparisons. Monte Carlo evidence provides some guidance as to when the two forms of analytics are most appropriate, in a nested model context.

JEL Nos.: C53, C52, C12
Keywords: predictability, forecast accuracy, testing

* *Clark*: Economic Research Dept.; Federal Reserve Bank of Kansas City; 1 Memorial Drive; Kansas City, MO 64198; todd.e.clark@kc.frb.org. *McCracken (corresponding author)*: Research Division.; Federal Reserve Bank of St. Louis; P.O. Box 442; St. Louis, MO 63166; michael.w.mccracken@stls.frb.org.
The authors gratefully acknowledge helpful comments from a reviewer and Michael Clements. 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, Federal Reserve Bank of St. Louis, or the Federal Reserve System.

1 Introduction

As noted throughout this Handbook, forecasting is a crucial part of economic and financial decision-making. The reason is clear: The efficacy of decisions made today depend on unknown future states of the world. Insofar as forecasts provide information about these unknown states, obtaining accurate forecasts is a critical step toward ensuring that good decisions are being made.

But how do we know that accurate forecasts are being made? Answering such a question first requires a definition of "accuracy." As a foil for discussion, suppose that at time T, we are interested in forecasting the value of a scalar y one period into the future. To do so, we use the information available to us, in the form of historical values of y_s, and other variables $x_s = (x_1, ..., x_k)'$, $s = 1, ..., T$, to form a prediction $\hat{y}_{T+1}$. When the future value of y eventually becomes known, accuracy can take many forms. The most common approach is to base this evaluation on how close $\hat{y}_{T+1}$ is to the realized value of y_{T+1} using a user-specific loss function $L(.)$ that depends on both $\hat{y}_{T+1}$ and y_{T+1}. The canonical example of such a loss function is quadratic and hence the numeric value of $(y_{T+1} - \hat{y}_{T+1})^2$ defines the degree of accuracy. Smaller values are more accurate than larger values. Other loss functions include absolute loss $|y_{T+1} - \hat{y}_{T+1}|$, linex $\exp(\alpha(y_{T+1} - \hat{y}_{T+1})) - \alpha(y_{T+1} - \hat{y}_{T+1}) - 1$, and so on.

Taken literally, at any given forecast origin T, such values are infeasible because there is simply no way to know the unknown future state of the world. A second-best approach to answering the original question is to ask a different one: On average, do we expect our forecasts to be accurate? That is, while we cannot possibly know whether the value of $(y_{T+1} - \hat{y}_{T+1})^2$ is small, can we know something about whether $E(y_{T+1} - \hat{y}_{T+1})^2$ is small, where $E(.)$ denotes the expectations operator? At face value the answer is typically "no" because we do not know the probability law governing the randomness associated with the unknown future (which, in turn, determines the magnitude of the forecast error $(y_{T+1} - \hat{y}_{T+1})$).

That said, under reasonable assumptions on the observables (the values of y and x known to the forecasting agent at time T), one can construct an estimate of $E(y_{T+1} - \hat{y}_{T+1})^2$ that does not require knowledge of y_{T+1} or the unknown expectations operator $E(.)$. The most common approach to doing this is to conduct what Stock and Watson (2003) refer to as a "pseudo-out-of-sample" forecasting exercise. The idea is pretty intuitive. Given

today's known observables $\{y_s, x_s\}_{s=1}^{T}$, define an "in-sample" portion of observables using observations $\{y_s, x_s\}_{s=1}^{R}$ and let the remaining observables $\{y_s, x_s\}_{s=R+1}^{T}$ define the "out-of-sample" portion. The thought experiment is to pretend that you are constructing forecasts $\hat{y}_{t+1}$ at each of the $P = T - R$ forecast origins $t = R, ..., T-1$ using the same methodology you propose to use at forecast origin T (but obviously with different information sets). At the end of this exercise, while still not knowing the true value of $E(y_{T+1} - \hat{y}_{T+1})^2$, we can construct $P^{-1}\sum_{t=R}^{T-1}(y_{t+1} - \hat{y}_{t+1})^2$ with the hope that this sample average provides some information about the unknown expected accuracy of the time T forecast.

In simplistic terms, tests of unconditional predictive ability are intended to answer this last, unknown aspect of the pseudo-out-of-sample exercise. Statistically, how close do we think $P^{-1}\sum_{t=R}^{T-1}(y_{t+1} - \hat{y}_{t+1})^2$ is to $E(y_{T+1} - \hat{y}_{T+1})^2$? More precisely, we can ask whether $P^{-1}\sum_{t=R}^{T-1}(y_{t+1} - \hat{y}_{t+1})^2$ is a consistent estimate of $E(y_{T+1} - \hat{y}_{T+1})^2$, ask whether we can form a meaningful confidence interval for $E(y_{T+1} - \hat{y}_{T+1})^2$, and more generally, ask whether we can conduct asymptotically valid inference about $E(y_{T+1} - \hat{y}_{T+1})^2$ based on $P^{-1}\sum_{t=R}^{T-1}(y_{t+1} - \hat{y}_{t+1})^2$. The answers to these questions — especially the last — depend on the type of data being used, the definition of accuracy, and the methods by which the forecasts are being constructed. Delineating these details is our prime objective.

Before doing so, it is worth noting how tests of unconditional predictive ability differ from tests of conditional predictive ability as discussed in Giacomini (this volume). Continuing with our previous example, the main difference involves asking questions not about $E(y_{T+1} - \hat{y}_{T+1})^2$ but rather about $E[(y_{T+1} - \hat{y}_{T+1})^2|\Im_T]$, where $E[\cdot|\Im_T]$ denotes the conditional expectation operator given an information set available to the forecasting agent at time T. At some level, tests of conditional predictive ability encompass those associated with tests of unconditional predictive ability. This arises because the user can always choose to be interested only in the trivial σ-field $(\varnothing, \Omega)$, in which case the conditional expectation operator is equivalent to the unconditional expectation operator. To date, as we detail in the body of this chapter, the unconditional and conditional testing literatures differ primarily in (1) the assumptions necessary to obtain asymptotic results and (2) some of the asymptotic results.

In the following sections we provide a brief overview of methods for constructing tests of unconditional predictive ability. In Section 2 we begin by describing the necessary notation for delineating these results. Sections 3 and 4 provide a discussion of the two main forms

of the null hypothesis: (i) population-level unconditional tests of predictive ability and (ii) finite-sample unconditional tests of predictive ability. For each we discuss the main methodological approaches to conducting inference, emphasizing results in West (1996), Clark and McCracken (2001, 2005b, 2009b), Giacomini and White (2006), and McCracken (2007).[1] In both instances, our discussion is in the context of using estimated parametric models to form a point prediction of a scalar dependent variable y. Other types of forecasts are referenced when relevant. Section 5 provides a brief overview of other developments in this literature and suggests new directions for future research. Section 6 provides Monte Carlo evidence on the efficacy of selected methods. Section 7 concludes.

2 The pseudo-out-of-sample environment

The sample of observations $\{y_t, x_t'\}_{t=1}^T$ includes a scalar random variable y_t to be predicted, as well as a $(k \times 1)$ vector of predictors x_t. Specifically, for each time t the variable to be predicted is $y_{t+\tau}$, where τ denotes the forecast horizon. The sample is divided into in-sample and out-of-sample portions. The total in-sample observations (on y_t and x_t) span 1 to R. Letting $P - \tau + 1$ denote the number of τ-step-ahead predictions, the total out-of-sample observations span $R + \tau$ through $R + P$. The total number of observations in the sample is $R + P = T$.[2]

The literature is largely silent on the best way to split the sample into in- and out-of-sample portions. There is, however, a clear trade-off. More out-of-sample observations (larger P) imply more forecasts and therefore more information regarding the accuracy of the forecasts. The converse is that more in-sample observations (larger R) imply that the parameter estimates will be more accurately estimated and likely lead to more accurate forecasts. As seen below, asymptotic inference regarding unconditional tests of predictive ability depends explicitly on the relative sample sizes P/R. This dependence, however, is largely a matter of asymptotic theory under the null hypothesis. To our knowledge, the only paper that considers the effect of the sample split under the alternative is Clark and McCracken (2005a), and even there attention is restricted to the effect on tests of predictive ability when there is unmodeled structural change.

[1] See West (2006) for a previous and detailed survey of the literature on the evaluation of population-level forecast accuracy.

[2] This seemingly innocuous assumption is actually nontrivial. For many macroeconomic variables (such as GDP) the forecasting agent actually has access to a triangular array of vintages of both the y's and x's. Such data are accessible at the Federal Reserve Banks of Philadelphia and St. Louis. Section 5.2 briefly discusses tests of unconditional predictive ability with real-time data.

In practice, the sample split (as measured by the ratio P/R) has taken a wide range of values, ranging from as small as 0.05 (Kuan and Liu, 1995) to as large as 10 (Breen et al., 1989). Equal splits are fairly common (in which case $P/R \approx 1$). One other common practice is to conduct two or more pseudo-out-of-sample exercises simultaneously as a means of observing whether the predictive ability has remained constant over time. For example, using data from 1958 to 2004 Clark and McCracken (2006) conduct two out-of-sample exercises, one in which the in-sample portion is 1958-1976 and the out-of-sample portion is 1977-1989, and another in which the in-sample portion is 1958-1989 and the out-of-sample portion is 1990-2003. By doing this, the accuracy of the models being considered can be observed over two distinct periods and any changes in performance can be detected — at least using an ocular test.

Given the sample split, forecasts of $y_{t+\tau}$, $t = R, \ldots, T-\tau$, are generated using parametric models of the form $y_{t+\tau} = g(x_t, \beta^*) + u_{t+\tau}$ for a known function $g(\cdot, \cdot)$ and unknown finite-dimensioned parameter β^*. These parameters are estimated using one of three distinct observation windows. Under the recursive scheme, the parameter vector is updated at each forecast origin $t = R, ..., T - \tau$ using all available information. For example, if NLLS is used to estimate the above model, we have $\hat{\beta}_t = \arg\min_\beta \sum_{s=1}^{t-\tau} (y_{s+\tau} - g(x_s, \beta))^2$. Under the rolling scheme, the parameters are also updated at each forecast origin but always using the same number of observations R in the window. Continuing with the same example this implies $\hat{\beta}_t = \arg\min_\beta \sum_{s=t-\tau-R+1}^{t-\tau} (y_{s+\tau} - g(x_s, \beta))^2$. In our final scheme — the fixed scheme — the parameters are estimated only once at the initial forecast origin and hence $\hat{\beta}_t = \hat{\beta}_R = \arg\min_\beta \sum_{s=1}^{R-\tau} (y_{s+\tau} - g(x_s, \beta))^2$.

Regardless of the sample window used, the parameter estimates and the predictors are used to construct forecasts $\hat{y}_{t+\tau}(x_t, \hat{\beta}_t) = \hat{y}_{t+\tau}$ of the dependent variable at each forecast origin. These in turn can be used to construct forecast errors $\hat{u}_{t+\tau} = y_{t+\tau} - \hat{y}_{t+\tau}$. Typically the accuracy of the forecasts is evaluated based on a known function of this forecast error. Table 1 provides a list of several of the most common measures of "accuracy," using our loose interpretation of the term. The first three measures are intended to evaluate the accuracy of a single model, whereas the remaining ones are better thought of as evaluating the accuracy of a model relative to another model.

Note that regardless of the measures of accuracy (from Table 1) of interest, each can be

Table 1. Common Measures of Forecast Accuracy

measure	$f_{t+\tau}(\beta)$
1. bias (zero mean prediction error)	$u_{t+\tau}$
2. serial correlation (zero first-order correlation)	$u_{t+\tau}u_{t+\tau-1}$
3. efficiency (no correlation between error and prediction)	$u_{t+\tau}g(x_t,\beta)$
4. encompassing (no correlation between model 1's error and model 2's prediction)	$u_{1,t+\tau}g_2(x_t,\beta)$
5. mean square error	$u^2_{t+\tau}$
6. mean absolute error	$\lvert u_{t+\tau}\rvert$
7. linex loss	$e^{\alpha u_{t+\tau}} - \alpha u_{t+\tau} - 1$

written in a general form as $f(y_{t+\tau}, x_t, \hat{\beta}_t) = f_{t+\tau}(\hat{\beta}_t)$.[3] Parallel to the example given in the introduction, the goal of tests of unconditional predictive ability is to determine how best to use $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ as a means of telling us something about the unknown future accuracy of the model(s). The following two subsections offer two distinct approaches to doing just that.

3 Population-Level Predictive Ability

A test of population-level predictive ability addresses the following question: Can we use $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ to learn something about $Ef_{t+\tau}(\beta^*)$? For this question, it is crucial that we recognize that $Ef_{t+\tau}(\beta^*)$ depends on β^*, the unknown true value of the parameter estimate $\hat{\beta}_t$. With this in mind, the original question can be recast as: Can $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ be used to learn something about the accuracy of the forecasts were we to know the true values of the model parameters?

3.1 West (1996)

Building on earlier work by Diebold and Mariano (1995), West (1996) develops a theory for addressing this question. In particular, he shows that

$$(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t) - Ef_{t+\tau}(\beta^*)) \rightarrow^d N(0,\Omega), \tag{1}$$

and hence for a given null hypothesis regarding $Ef_{t+\tau}(\beta^*)$, asymptotically valid inference can be conducted using standard normal critical values so long as one can obtain an asymptotically valid estimate of Ω.[4]

[3] When two models are involved, redefine $\hat{\beta}_t$ as the vector formed by stacking the parameter estimates from each of the two models so that $\hat{\beta}_t = (\hat{\beta}'_{1,t}, \hat{\beta}'_{2,t})'$.

[4] Studies such as Corradi and Swanson (2007) have developed bootstrap-based inference approaches that can be applied with tests that have power against generic alternatives or with tests applied to forecasts from misspecified models.

The details of how to estimate Ω is perhaps the main technical development in West (1996). Before providing this result, some notation and assumptions are needed.[5]

(A1) $\hat{\beta}_t = \beta^* + BH(t) + o_{a.s.}(1)$, where for some mean zero process $h_{t+\tau} = h_{t+\tau}(\beta^*)$ [with h denoting the orthogonality conditions used to estimate parameters, such as $h_{t+\tau} = x_t u_{t+\tau}$ for a single linear regression], $H(t)$ equals $t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau}$, $R^{-1}\sum_{s=t-R+1}^{t-\tau} h_{s+\tau}$, and $R^{-1}\sum_{s=1}^{R-\tau} h_{s+\tau}$ for the recursive, rolling, and fixed schemes, respectively, and B denotes a nonstochastic matrix.

(A2) The vector $(f_{t+\tau}(\beta^*), h'_{t+\tau})'$ is covariance stationary and satisfies mild mixing and moment conditions.

(A3) $\lim_{P,R\to\infty} P/R \to \pi$, a constant that is finite for the rolling and fixed schemes but can be infinite for the recursive scheme.

(A4) The vector $F = E[\partial f_{t+\tau}(\beta)/\partial\beta]_{\beta=\beta^*}$ is finite.[6]

(A5) Ω is positive.

Given these assumptions, West (1996) shows that the asymptotic variance Ω can take a variety of forms depending on how the parameters are estimated:

$$\Omega = S_{ff} + \lambda_{fh}(FBS'_{fh} + S_{fh}B'F') + \lambda_{hh}FBS_{hh}B'F', \tag{2}$$

where $S_{ff} = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} f_{t+\tau}(\beta^*))$, $S_{hh} = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} h_{t+\tau})$, $S_{fh} = \lim_{T\to\infty} Cov(T^{-1/2}\sum_{s=1}^{T-\tau} f_{t+\tau}(\beta^*), T^{-1/2}\sum_{s=1}^{T-\tau} h_{t+\tau})$, and

	$\lambda_{fh} =$	$\lambda_{hh} =$
Recursive	$1 - \pi^{-1}\ln(1+\pi)$	$2(1 - \pi^{-1}\ln(1+\pi))$
Rolling, $\pi \leq 1$	$\pi/2$	$\pi - \pi^2/3$
Rolling, $1 < \pi < \infty$	$1 - (2\pi)^{-1}$	$1 - (3\pi)^{-1}$
Fixed	0	π

.

In equation (2) we see that Ω consists of three terms. The first, S_{ff}, is the long-run variance of the measure of accuracy when the parameters are known. The third term, $\lambda_{hh}FBS_{hh}B'F'$, captures the contribution of the variance due purely to the fact that we do not observe β^* but must estimate it instead. The second term, $\lambda_{fh}(FBS'_{fh} + S_{fh}B'F')$, captures the covariance between the measure of accuracy and the estimation error associated with $\hat{\beta}_t$. Because the parameter estimates can be constructed using three different observation windows (recursive, rolling, and fixed) it is not surprising that the terms that arise due to estimation error depend on that choice via the terms λ_{fh} and λ_{hh}.

[5] These assumptions are intended to be expository, not complete. See West (1996) for more detail.

[6] McCracken (2000) weakens this assumption to $F = \partial E[f_{t+\tau}(\beta)]/\partial\beta_{\beta=\beta^*}$ so that the function $f_{t+\tau}(\beta)$ need not be differentiable.

With this formula in hand, estimating Ω is straightforward. Since $\hat{\pi} = P/R \to \pi$ and both λ_{fh} and λ_{hh} are continuous in π, substituting $\hat{\pi}$ for π is sufficient for estimating both λ_{fh} and λ_{hh}. The F term can be estimated directly using $\hat{F} = (P - \tau + 1)^{-1} \sum_{t=R}^{T-\tau} \partial f_{t+\tau}(\hat{\beta}_t)/\partial\beta$. When only one model has been estimated, the B term is typically the inverse of the Hessian matrix associated with the loss function used to estimate the model parameters. For example, if NLLS is used to estimate the model such that $\hat{\beta}_t = \arg\min_\beta \sum_{s=1}^{t-\tau} (y_{s+\tau} - g(x_s, \beta))^2$, then a consistent estimate of B is given by $\hat{B} = (T^{-1} \sum_{s=1}^{T-\tau} \partial^2 (y_{s+\tau} - g(x_s, \hat{\beta}_T))^2/\partial\beta\partial\beta')^{-1}$. If more than one model is being used to construct $f_{t+\tau}(\hat{\beta}_t)$ (so that $\hat{\beta}_t = (\hat{\beta}'_{1,t}, \hat{\beta}'_{2,t})'$), then B is the block diagonal matrix $diag(B_1, B_2)$ and hence a consistent estimate is $\hat{B} = diag(\hat{B}_1, \hat{B}_2)$.

For the long-run variances and covariances, West (1996) shows that standard kernel-based estimators are consistent. To be more precise, define $\bar{f} = (P-\tau+1)^{-1} \sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$, $\hat{\Gamma}_{ff}(j) = (P-\tau+1)^{-1} \sum_{t=R+j}^{T-\tau} (f_{t+\tau}(\hat{\beta}_t) - \bar{f})(f_{t+\tau-j}(\hat{\beta}_{t-j}) - \bar{f})$, $\hat{\Gamma}_{hh}(j) = T^{-1} \sum_{t=j+1}^{T-\tau} h_{t+\tau}(\hat{\beta}_t) \times h'_{t+\tau-j}(\hat{\beta}_{t-j})$ and $\hat{\Gamma}_{fh}(j) = (P - \tau + 1)^{-1} \sum_{t=R+j}^{T-\tau} f_{t+\tau}(\hat{\beta}_t) h'_{t+\tau-j}(\hat{\beta}_{t-j})$, with $\hat{\Gamma}_{ff}(j) = \hat{\Gamma}_{ff}(-j)$, $\hat{\Gamma}_{hh}(j) = \hat{\Gamma}'_{hh}(-j)$, and $\hat{\Gamma}_{fh}(j) = \hat{\Gamma}'_{fh}(-j)$. The long-run variance estimates $\hat{S}_{ff}$, $\hat{S}_{hh}$, and $\hat{S}_{fh}$ are then constructed, as in Newey and West's (1987) HAC estimator, by weighting the relevant leads and lags of these covariances.

Interestingly, for some cases estimating Ω is as simple as using the estimate $\hat{\Omega} = \hat{S}_{ff}$. This arises when the second and third terms in equation (2), those due to estimation error, cancel and hence we say the estimation error is asymptotically irrelevant.

Case 1. If $\pi = 0$, then both λ_{fh} and λ_{hh} are zero and hence $\Omega = S_{ff}$. This case arises naturally when the sample split is chosen so that the number of out-of-sample observations is small relative to the number of in-sample observations.[7]

Case 2. If $F = 0$, then $\Omega = S_{ff}$. This case arises under certain very specific circumstances but arises most naturally when the measure of "accuracy" is explicitly used when estimating the model parameters. The canonical example is the use of a quadratic loss function to evaluate the accuracy of forecasts from two non-nested models estimated by NLLS. In this situation, the F term equals zero and estimation error is asymptotically irrelevant.

Case 3. Under the recursive scheme, there are instances where $-S_{fh}B'F' = FBS_{hh}B'F'$. In this case, it isn't so much than any particular term equals zero but that the sum of the

[7]Chong and Hendry (1986) first observed that the parameter estimation error is irrelevant if P is small relative to R.

components just happens to cancel to zero. See West (1996) and West and McCracken (1998) for some examples.

3.2 Clark and McCracken (2001, 2005b), McCracken (2007)

Although the results in West (1996) have many applications, the theory is not universal. In particular, one of the primary assumptions for the results in West (1996) to hold is that Ω must be positive. In nearly all the examples from Table 1, this is not an issue. However, problems arise in applications where one wishes to compare the accuracy of two models — but the models are nested under the null of equal unconditional forecast accuracy. Consider the case where two nested OLS-estimated linear models are being compared. If we define the $(k \times 1, k = k_1 + k_2)$ vector of predictors $x_t = x_{2,t} = (x'_{1,t}, x'_{12,t})'$, the models take the form $y_{t+\tau} = x'_{i,t}\beta^*_i + u_{i,t+\tau}$, for $i = 1, 2$, such that model 2 nests model 1 and hence $\beta^*_2 = (\beta^{*\prime}_1, \beta^{*\prime}_{12})' = (\beta^{*\prime}_1, 0)'$ under the null. If we use quadratic loss to measure accuracy, we find that $f_{t+\tau}(\beta^*) = (y_{t+\tau} - x'_{1,t}\beta^*_1)^2 - (y_{t+\tau} - x'_{2,t}\beta^*_2)^2 = (y_{t+\tau} - x'_{1,t}\beta^*_1)^2 - (y_{t+\tau} - x'_{1,t}\beta^*_1)^2 = 0$ for all t, and hence it is clearly the case that S_{ff}, S_{fh}, and F all equal zero, making Ω also equal zero.

In this case, Clark and McCracken (2001, 2005b) and McCracken (2007) develop a different set of asymptotics that allow for an out-of-sample test of equal population-level unconditional predictive ability between two nested models. The key to their theory is to note that while $P^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t) - 0) \to^p 0$ when the models are nested, $\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t) - 0)$ need not have a degenerate asymptotic distribution. Building on this insight they show that, in the context of linear, OLS-estimated, direct-multistep forecasting models, a variety of statistics can be used to test for equal forecast accuracy and forecast encompassing despite the fact that the models are nested. Let $\hat{u}_{i,t+\tau} = y_{t+\tau} - x'_{i,t}\hat{\beta}_{i,t}$, $i = 1, 2$, $\hat{d}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}$, $\hat{c}_{t+\tau} = \hat{u}_{1,t+\tau}(\hat{u}_{1,t+\tau} - \hat{u}_{2,t+\tau})$, and $\hat{\sigma}^2_2 = (P - \tau + 1)^{-1}\sum_{t=R}^{T-\tau}\hat{u}^2_{2,t+\tau}$. If we let $\hat{S}_{dd}$ and $\hat{S}_{cc}$ denote long-run variance estimates for, respectively, $\hat{d}_{t+\tau}$ and $\hat{c}_{t+\tau}$ (analogous to $\hat{S}_{ff}$ above) constructed as in Newey and West's (1987) HAC estimator, these statistics take the form

$$\text{MSE-}t = \frac{P^{-1/2}\sum_{t=R}^{T-\tau}\hat{d}_{t+\tau}}{\hat{S}^{1/2}_{dd}}, \quad \text{MSE-}F = \frac{\sum_{t=R}^{T-\tau}\hat{d}_{t+\tau}}{\hat{\sigma}^2_2} \tag{3}$$

$$\text{ENC-}t = \frac{P^{-1/2}\sum_{t=R}^{T-\tau}\hat{c}_{t+\tau}}{\hat{S}^{1/2}_{cc}}, \quad \text{ENC-}F = \frac{\sum_{t=R}^{T-\tau}\hat{c}_{t+\tau}}{\hat{\sigma}^2_2}. \tag{4}$$

In each case the asymptotic distributions have representations as functions of stochastic integrals of quadratics in Brownian motion. In limited cases, where the models are correctly specified and the forecast errors are serially uncorrelated and exhibit conditional homoskedasticity, critical values are tabulated.[8]

For all other cases, the asymptotic distributions depend on unknown nuisance parameters that capture the presence of serial correlation in the forecast errors and conditional heteroskedasticity. Since tabulating critical values in the general case is infeasible, in the following we present a simple bootstrap developed in Clark and McCracken (2009b) that provides asymptotically valid critical values.

1. (a) Use OLS to estimate the parameter vector β_1^* associated with the restricted model. Store the fitted values $x'_{1,s}\hat{\beta}_{1,T}$, $s = 1, ..., T$. (b) Use OLS to estimate the parameter vector β_2^* associated with the unrestricted model. Store the residuals $\hat{v}_{2,s+\tau}$, $s = 1, ..., T$.

2. Using NLLS, estimate an $MA(\tau - 1)$ model for the OLS residuals $\widehat{v}_{2,s+\tau}$ such that $v_{2,s+\tau} = \varepsilon_{2,s+\tau} + \theta_1\varepsilon_{2,s+\tau-1} + ... + \theta_{\tau-1}\varepsilon_{2,s+1}$. Let $\eta_{s+\tau}$, $s = 1, ..., T$, denote an $i.i.d$ $N(0, 1)$ sequence of simulated random variables. Define $\widehat{v}^*_{2,s+\tau} = (\eta_{s+\tau}\widehat{\varepsilon}_{2,s+\tau} + \widehat{\theta}_1\eta_{s-1+\tau}\widehat{\varepsilon}_{2,s+\tau-1} + ... + \widehat{\theta}_{\tau-1}\eta_{s+1}\widehat{\varepsilon}_{2,s+1})$, $s = 1, ..., T$. Form artificial samples of $y^*_{s+\tau}$ using the fixed regressor structure, $y^*_{s+\tau} = x'_{1,s}\hat{\beta}_{1,T} + \widehat{v}^*_{2,s+\tau}$.

3. Using the artificial data, construct an estimate of the test statistics (e.g., MSE-F, MSE-t, ENC-F, ENC-t) as if these were the original data.

4. Repeat steps 2 and 3 a large number of times: $j = 1, ..., N$.

5. Reject the null hypothesis, at the α% level, if the test statistic is greater than the $(100 - \alpha)$%-ile of the empirical distribution of the simulated test statistics.

Both analytically and via a range of Monte Carlo simulations, Clark and McCracken (2009b) show that this bootstrap provides valid — asymptotically and in practice — critical values for testing the null of equal population-level predictive ability between two nested models. Further evidence is given in Section 6.

4 Finite-Sample Predictive Ability

A test of finite-sample predictive ability addresses a different, but related, question than the one described in the previous section: Can we use $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ to learn something about $Ef_{t+\tau}(\hat{\beta}_t)$? For this question, it is crucial to recognize that $Ef_{t+\tau}(\hat{\beta}_t)$

[8] We have made available at www.estima.com a computer procedure in the RATS software package that generates 1-step ahead forecasts, test statistics, and asymptotic critical values.

depends on $\hat{\beta}_t$ and not the unknown true value of the parameter β^*. In other words, we want to know whether $(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} f_{t+\tau}(\hat{\beta}_t)$ can be used to learn something about the accuracy of the forecasts given that our forecasts are constructed using estimated parameters.

The importance of such a distinction is perhaps easiest to see when comparing the forecast accuracy of two nested models. Continuing with the notation above, suppose the two models take the form $y_{t+\tau} = x_{i,t}'\beta_i^* + u_{i,t+\tau}$ for $i = 1, 2$. We know that if $\beta_{12}^* = 0$, then the two models are identical and hence have equal population-level predictive ability. We also know that if $\beta_{12}^* \neq 0$, then in population, the larger model will forecast more accurately than the smaller model. In practice, though, even when $\beta_{12}^* \neq 0$, the parameters are estimated with finite samples of data. It is then perfectly reasonable to consider the option that the smaller model is as accurate as (or even more accurate than) the larger model despite the fact that $\beta_{12}^* \neq 0$. This is particularly likely when the dimension of β_{12}^* is large relative to the existing sample size.

4.1 Giacomini and White (2006)

The first study to address this type of null hypothesis is Giacomini and White (2006). They note that two models can have equal forecast accuracy in finite samples if, continuing with our nested model comparison, the bias associated with estimating the misspecified restricted model happens to balance with the additional estimation error associated with estimating β_{12}^* in the correctly specified unrestricted model. This observation is perfectly true, but implementing a test for it is much harder, especially given a universe where you don't want to have to make extremely restrictive assumptions on the data (such as joint normality, conditionally homoskedastic and serially uncorrelated forecast errors, etc.). This scenario is much harder because we know in advance that any asymptotic approach to inference that allows the parameter estimates to be consistent for their population counterparts will imply that the unrestricted model is more accurate than the restricted model. In the notation of the tests of population-level predictive ability and our nested model comparison, this implies that any asymptotics that allow R to diverge to infinity will fail to be relevant for the null of equal finite-sample unconditional predictive ability.

As a result, Giacomini and White (2006) dispense with that assumption. More precisely they show that if the parameter estimates are constructed using a *rolling scheme with a*

finite observation window R, then

$$(P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-Ef_{t+\tau}(\hat{\beta}_t)) \rightarrow^d N(0,V), \tag{5}$$

where $V = S_{\hat{f}\hat{f}} = \lim_{P\rightarrow\infty} Var((P-\tau+1)^{-1/2}\sum_{t=R}^{T-\tau}(f_{t+\tau}(\hat{\beta}_t)-Ef_{t+\tau}(\hat{\beta}_t)))$. Note that this differs from the asymptotic variance in West (1996) even when the second and third terms in Ω are asymptotically irrelevant since $S_{\hat{f}\hat{f}} \neq S_{ff}$.

This result is extremely powerful and covers a wide range of applications, including every example in Table 1. Interestingly, by requiring that the forecasts be constructed using a small, finite, rolling window of observations, Giacomini and White (2006) are able to substantially weaken many of the most important assumptions needed for the results in West (1996) and Clark and McCracken (2001, 2005b). In particular, covariance stationarity of the observables is no longer needed — only that the observables are $I(0)$ with relatively mild mixing and moment conditions. There is no need for Ω to be positive (though V must be), and hence both nested and non-nested comparisons are allowed.[9]

The primary weakness of the results in Giacomini and White (2006) is that their approach cannot be used with the recursive scheme. The recursive scheme fails because, absent any other assumptions on the parameter β_{12}^*, as the sample size increases the parameter estimates $\hat{\beta}_t$ are consistent for their population counterparts and thus estimation error vanishes. Although the rolling scheme is relatively common among forecasting agents, it is by no means universal. Moreover, the asymptotics apply only when we think of the rolling observation window as small relative to the number of out-of-sample observations. Monte Carlo evidence on the magnitudes of P and R needed for accurate inference is limited.

4.2 Clark and McCracken (2009b)

More recent work by Clark and McCracken (2009b) shows that, in some circumstances, one can construct a test of equal finite-sample unconditional predictive ability that permits not only the rolling scheme, but also the recursive scheme. In particular, they consider the case of testing this null hypothesis when comparing two OLS-estimated linear models and hence $Ef_{t+\tau}(\hat{\beta}_t) = E[(y_{t+\tau} - x_{1,t}'\hat{\beta}_{1,t})^2 - (y_{t+\tau} - x_{2,t}'\hat{\beta}_{2,t})^2] = 0$. The asymptotics are not unlike those from their previous work on equal population-level predictive ability (described in the previous section) but capture the bias and estimation error associated

[9] While not the focus of this review, the forecasts can be based on estimators that are Bayesian, nonparametric, or semi-parametric. The key is that R must be small and finite in all cases.

with, respectively, a misspecified restricted model and a correctly specified, but imprecisely estimated, unrestricted model.

But as noted above, since their results are asymptotic and the estimation error associated with the parameter estimates vanishes asymptotically, balancing that estimation error with a bias component is problematic using standard parameterizations of a linear regression model. Instead Clark and McCracken (2009b) consider the case in which the additional predictors in the unrestricted model are "weak" using the following local-to-zero parameterization:

$$y_{t+\tau} = x'_{2,t}\beta^*_{2,R} + u_{t+\tau} = x'_{1,t}\beta^*_1 + x'_{12,t}(R^{-1/2}\beta^*_{12}) + u_{t+\tau}. \tag{6}$$

The intuition for this parameterization is based on an observation: As the sample size used to estimate the regression parameters increases, the estimation error vanishes at a root-R rate. If bias due to model misspecification in the smaller (restricted) model is going to balance with the estimation error, it must also vanish at a root-R rate. To be clear, we do not take the model in equation (6) as a literal representation of the data, but rather consider it a tool for modeling how a bias-variance trade-off can exist in large samples as the size of the sample used for estimation increases.

With this parameterization, on average over the out-of-sample period, the two models have equal unconditional forecast accuracy if for $B_i = (Ex_{i,t}x'_{i,t})^{-1}$, $i = 1, 2$, $J_2 = (0_{k_2 \times k_1}, I_{k_2 \times k_2})'$, $F_2 = J'_2 B_2 J_2$, and $V = \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T-\tau} h_{2,s+\tau})$,

$$\frac{\beta^{*\prime}_{22} F_2^{-1} \beta^*_{22}}{tr((-JB_1J' + B_2)V)} = \delta,$$

where δ takes the value $\frac{\ln(1+\pi)}{\pi}$ and 1 for the recursive and rolling schemes, respectively. In the case of one-step-ahead forecasts with conditionally homoskedastic errors, this simplifies since $tr((-JB_1J' + B_2)V) = \sigma^2 k_2$.

However, as was the case for tests of equal population-level forecast accuracy between two nested models, the asymptotic distributions derived by Clark and McCracken (2009b) are nonstandard and have representations as functions of stochastic integrals of quadratics in Brownian motion. Moreover, the asymptotic distributions depend on unknown nuisance parameters that capture the presence of serial correlation in the forecast errors and conditional heteroskedasticity. Since tabulating critical values in the general case is infeasible, in the following we present a simple bootstrap (not unlike the one presented previously) that

can provide asymptotically valid critical values in certain circumstances. In the following, let $\hat{\pi} = P/R$, $B_i(T) = (T^{-1}\sum_{s=1}^{T-\tau} x_{i,s}x'_{i,s})^{-1}$, and let $V(T)$ denote a HAC estimator of the long-run variance of the OLS moment condition $\hat{v}_{2,s+\tau}x_{2,s}$ associated with the unrestricted model.

1. (a) Estimate the parameter vector β_2^* associated with the unrestricted model using the weighted ridge regression

$$\begin{aligned} \tilde{\beta}_{2,T} &= (\tilde{\beta}'_{12,T}, \tilde{\beta}'_{22,T})' \\ &= \arg\min_{b_2} \sum_{s=1}^{T-\tau} (y_{s+\tau} - x'_{2,s}b_2)^2 \text{ s.t. } b_2'J_2F_2^{-1}(T)J_2'b_2 = \hat{\delta}/R, \end{aligned} \tag{7}$$

where $\hat{\delta}$ equals $\frac{\ln(1+\hat{\pi})}{\hat{\pi}} tr((-JB_1(T)J' + B_2(T))V(T))$ or $tr((-JB_1(T)J' + B_2(T))V(T))$ for the recursive or rolling schemes, respectively. Store the fitted values $x'_{2,t}\tilde{\beta}_{2,T}$. (b) Estimate the parameter vector β_2^* associated with the unrestricted model using OLS and store the residuals $\hat{v}_{2,s+\tau}$.

2. Using NLLS, estimate an $MA(\tau-1)$ model for the OLS residuals $\hat{v}_{2,s+\tau}$ such that $v_{2,s+\tau} = \varepsilon_{2,s+\tau} + \theta_1\varepsilon_{2,s+\tau-1} + ... + \theta_{\tau-1}\varepsilon_{2,s+1}$. Let $\eta_{s+\tau}$, $s = 1, ..., T$, denote an $i.i.d$ $N(0,1)$ sequence of simulated random variables. Define $\hat{v}^*_{2,s+\tau} = (\eta_{s+\tau}\hat{\varepsilon}_{2,s+\tau} + \hat{\theta}_1\eta_{s-1+\tau}\hat{\varepsilon}_{2,s+\tau-1} + ... + \hat{\theta}_{\tau-1}\eta_{s+1}\hat{\varepsilon}_{2,s+1})$, $s = 1, ..., T$. Form artificial samples of $y^*_{s+\tau}$ using the fixed regressor structure, $y^*_{s+\tau} = x'_{2,s}\tilde{\beta}_{2,T} + \hat{v}^*_{2,s+\tau}$.

3. Using the artificial data, construct an estimate of the test statistics (e.g., MSE-F, MSE-t) as if this were the original data.

4. Repeat steps 2 and 3 a large number of times: $j = 1, ..., N$.

5. Reject the null hypothesis, at the α% level, if the test statistic is greater than the $(100-\alpha)$%-ile of the empirical distribution of the simulated test statistics.

Clark and McCracken (2009b) show that critical values from this bootstrap are asymptotically valid in two important special cases. First, if the number of additional predictors (k_2) is 1, then the bootstrap is asymptotically valid and allows for both multiple-step-ahead forecasts and conditionally heteroskedastic errors. Second, if the forecast horizon (τ) is 1 and the forecast errors are conditionally homoskedastic, then the bootstrap is asymptotically valid even when the number of additional predictors is greater than 1. But in the most general case where $k_2 > 1$ and either the forecast horizon is greater than 1 or the forecast errors are conditionally heteroskedastic, then the bootstrap is not asymptotically valid. Regardless, Monte Carlo evidence in Clark and McCracken (2009b) suggests that

even when the bootstrap validity conditions fail, the size distortions are not particularly severe.

5 Other Dimensions of Predictive Ability

This section briefly reviews existing extensions of the research previously cited on tests for equal unconditional predictive ability, extensions to real-time data, and some areas for future research.

5.1 Existing extensions

The research described above establishing the properties of important tests of equal predictive ability generally applies only under some limitations, including covariance stationarity of the data and an interest in point forecasts from a single model or pair of models. Other researchers have developed testing methods that address these limitations. For example, Corradi, Swanson, and Olivetti (2001) extend the results of West (1996), Clark and McCracken (2001, 2005b), and McCracken (2007) on tests of equal MSE with non-nested and nested models to allow for data with unit roots and cointegrating relationships. Their conditional moment-type test of predictive accuracy can be compared against critical values computed with a conditional p-value approach. Rossi (2005) extends prior theory to tackle the evaluation of long-horizon forecasts generated with persistent regressors.

Other researchers have developed tests of superior predictive ability for use with multiple models. White (2000) uses the asymptotics of West (1996) to develop a bootstrap-based approach to determining whether the best forecast among many is significantly more accurate. Hansen's (2005) modified version of White's test has better power and is less sensitive to especially poor alternative forecasts. The dependence of the White and Hansen asymptotics on West (1996) means that these tests are appropriately applied with forecasts from non-nested models, but not nested models. Hubrich and West (2009) develop some approaches to testing for equal accuracy at the population level in a small set of nested models. One such approach involves adjusting t-tests for equal MSE for the impact of parameter estimation error and comparing the maximum test statistic against critical values obtained from simulations of the maximum of a set of normal random variables.

Finally, other studies have extended existing work on point forecasts by developing tests for density forecasts. Although most of the density evaluation literature has abstracted

from forecasts from estimated models (e.g., Christoffersen, 1998), some recent work has developed testing methods applicable to forecasts from estimated models. Using West (1996)-type asymptotics, Corradi and Swanson (2006) use bootstrap methods to develop a density accuracy test that can be applied to forecasts from non-nested models. Amisano and Giacomini (2007) use the asymptotics of Giacomini and White (2006) to develop a density test — in its simplest form, a t-test for the equality of the log predictive score — that can be applied to forecasts from nested or non-nested models, as long as the models are estimated with a rolling sample of data.

5.2 Real-time data

The aforementioned literature on tests of equal predictive ability has generally ignored the real-time nature of the data used in many applications. Specifically, most research has abstracted from the possibility that, at any given forecast origin, the most recent data is subject to revision. While a data revision process is very difficult to accommodate in asymptotic theory for forecast tests, out-of-sample tests of equal predictive ability may be affected significantly by changes in the correlation structure of the data as the revision process unfolds. This susceptibility to data revisions has three sources: (i) while parameter estimates are typically functions of only a small number of observations that remain subject to revision, out-of-sample statistics are functions of a sequence of parameter estimates (one for each forecast origin $t = R, \ldots, T - \tau$), (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 data subject to revision possess a different mean and covariance structure than final revised data, tests of predictive ability using real-time data may have a different asymptotic distribution than tests constructed using data that is never revised.

To our knowledge, as of this writing, Clark and McCracken (2009a) represents the only existing extension of tests of equal predictive ability to the case of data subject to revision. The results of that study indicate data revisions can significantly affect the asymptotic behavior of tests of equal population-level predictive ability. For example, for tests applied to forecasts from non-nested models, West (1996) shows 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 when revisions are news, as defined in Mankiw, Runkle and Shapiro (1984).

When even some noise is present, parameter estimation error contributes to the asymptotic variance of the test statistic and cannot be ignored in inference. As another example, for nested model tests of equal predictive ability, Clark and McCracken (2001, 2005b) and McCracken (2007) show 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, Clark and McCracken (2009a) show that the standard test statistics diverge with probability one under the null hypothesis. They introduce a variant of the standard test statistic that is asymptotically standard normal despite being a comparison between two nested models.

Much remains to be done to extend out-of-sample inference methods to accommodate real-time data. One particular challenge is to develop methods for testing equal finite-sample predictive ability in forecasts from real-time data. Another is to develop tests of density forecasts made with real-time data.

5.3 Topics for future research

Beyond forecast inference with real-time data, there remain many other important topics for future research. One such topic is the development of tests of predictive ability based on forecasts from nested models estimated with data incorporating unit roots and cointegrating relationships. It would also be useful to extend the results of Clark and McCracken (2009b) on finite-sample predictive ability from their case of nested models to the case of non-nested models. Another important topic is the optimal choice of the sample split — that is, how best to divide a given sample between observations for initial estimation and for forecast evaluation. To this point, there has been no analytical work on this important practical question. Yet another area for further research would be predictive ability tests for models that overlap in the sense of Vuong (1989) — the two competing forecast models are not nested with each other but nest a restricted, baseline model that is, in fact, the data-generating process.

Perhaps the most glaring omission in this literature is simply a complete characterization of why one would use out-of-sample methods in the first place. Although arguments in favor of out-of-sample methods often reference robustness to data mining (e.g., Ashley et al., 1980), Inoue and Kilian (2004) argue that out-of-sample methods are no more robust to data-mining than in-sample methods provided appropriate critical values are used. They further argue, both analytically and via Monte Carlo simulations, that out-of-sample tests

of predictive ability have lower power than their in-sample counterparts. In our opinion, the strongest argument in favor of out-of-sample methods is made (albeit opaquely) in Clark and McCracken (2005a). There they show that the sequential nature of out-of-sample tests allows them to detect not only predictive ability, but also *changes* in predictive ability due to unmodeled structural change that in-sample tests cannot detect by construction.[10]

6 Monte Carlo Evidence

In the interest of brevity, we focus our Monte Carlo investigation on a limited subset of the many issues with testing for unconditional predictive ability, with an eye toward current, relatively unexplored topics. Specifically, our experiments focus on the following issues in testing for equal MSE in forecasts from nested models: (1) equal accuracy in the population versus the finite sample and (2) inference based on the asymptotics of Giacomini and White (2006) versus Clark and McCracken (2009b). Accordingly, our simulation results are mostly closed related to those of Giacomini and White (2006) and Clark and McCracken (2009b).

These experiments are based on data-generating processes (DGPs) from common macroeconomic applications. In all cases, the benchmark forecasting model is a univariate model of the predictand y; the alternative models add lags of various other variables of interest. The general null hypothesis is that the forecast from the alternative model is no more accurate than the benchmark forecast. This general null, however, takes different specific forms: (1) The models' forecasts are equally accurate in population, due to the variables in the alternative model having no predictive content (coefficients of 0); or (2) the models' forecasts are equally accurate in the finite sample, because the relevant coefficients in the alternative model are nonzero but small.

To distinguish the results in this paper, we depart from the focus of our prior papers in some important ways. First, in this paper, we focus on forecasts generated under the rolling scheme and the MSE-t test for equal accuracy. Under a rolling scheme, both the bootstrap method developed in Clark and McCracken (2009b) and standard normal inference as in Giacomini and White (2006) are asymptotically valid for inference with the MSE-t test. In contrast, in our prior work (such as Clark and McCracken, 2009b), we focused on recursive forecasts and presented only a limited set of results under the rolling scheme. In practice, though, the distinction between recursive and rolling forecasts is modest; the results we

[10] Giacomini and Rossi (2009) develop methods of testing for forecast breakdowns based on the asymptotics of Giacomini and White (2006).

present here also apply in recursive test results not reported. Second, we consider a larger and wider array of sample sizes than in our past work. We consider P/R ratios of 0.1, 0.2, 0.5, 1, 2, and 3, with R set at 50, 100, 150, and 200. Finally, as detailed below, for experiments under the null of equal accuracy in a finite sample, we refine the Monte Carlo settings of Clark and McCracken (2009b) by using preliminary sets of simulations to adjust the coefficients of the DGPs to be sure the competing models' MSEs can be expected to be equal on average.

6.1 Monte Carlo design

The Monte Carlo design is largely the same as that used by Clark and McCracken (2009b). The DGPs are based on empirical relationships among U.S. inflation and a range of predictors, estimated with 1968-2008 data. In the interest of brevity, we focus on a forecast horizon of one step. For all DGPs, we generate data using independent draws of innovations from the normal distribution and the autoregressive structure of the DGP. In all cases, our reported results are based on 5000 Monte Carlo draws and 499 bootstrap replications.

6.1.1 DGPs

DGP 1 is based on the empirical relationship between the change in core inflation (y_t) and the Chicago Fed's National Activity Index of the business cycle ($x_{1,t}$, the CFNAI):

$$\begin{aligned} y_{t+1} &= -0.4y_t - 0.1y_{t-1} + b_{11}x_{1,t} + u_{t+1} \\ x_{1,t+1} &= 0.7x_{1,t} + v_{1,t+1} \\ & \quad \text{var}\begin{pmatrix} u_{t+1} \\ v_{1,t+1} \end{pmatrix} = \begin{pmatrix} 0.8 & \\ 0.0 & 0.3 \end{pmatrix}. \end{aligned} \tag{8}$$

In the DGP 1 experiments, the alternative (unrestricted) forecasting model takes the form of the DGP equation for y_{t+1} (with constant added); the null or benchmark (restricted) model drops $x_{1,t}$:

$$\text{null: } y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + u_{0,t+1}. \tag{9}$$

$$\text{alternative: } y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 x_{1,t} + u_{1,t+1}. \tag{10}$$

We consider various experiments with different settings of b_{11}, the coefficient on $x_{1,t}$, which corresponds to the elements of our theoretical construct $\beta^*_{12}/\sqrt{R}$ in (6). In one set of simulations (Table 2), the coefficient is set to 0 to make the competing forecasting models equally accurate in population. In another set of experiments (Table 5), the coefficient is

set to 0.3, such that the alternative model is expected to be more accurate than the null, in population and in the finite sample.

In experiments under the null of equal accuracy in the finite sample (Tables 3 and 4), the coefficient is set to a value that makes the models equally accurate (in expectation) on average over the forecast sample. We used the asymptotic approximations developed in Clark and McCracken (2009b) and Monte Carlo experiments to set the coefficient value for each R, P combination. Specifically, for a given R, P setting, we first use the asymptotics to determine an equal-accuracy coefficient value. For example, with $R = P = 100$, the asymptotics yield a coefficient value of $b_{11} = 0.1166$, about one-half of the empirical estimate. We then conduct two sets of Monte Carlo experiments with a large number of draws, searching across grids of coefficient values — centered around the asymptotically-set value — to select a coefficient value that minimizes the average (across Monte Carlo draws) difference in MSEs from the competing forecasting models.[11] This Monte Carlo search typically results in some small upward adjustment of the asymptotically set coefficient. In the experiment with $R = P = 100$, the Monte Carlo refinement yields a coefficient setting of $b_{11} = 0.1245$, which is the value used to generate the DGP 1 results given in Tables 3 and 4 for $R = P = 100$. Overall, this Monte Carlo refinement yields models that appear to truly be equally accurate in a finite sample, with average MSE differences (Model 1 less Model 2) that are at most 0.0001 percentage point in absolute value. By comparison, the population error variance for the DGP equation is 0.8.

DGP 2 is based on the empirical relationship of the change in core inflation (y_t) to the CFNAI ($x_{1,t}$), food price inflation less core inflation ($x_{2,t}$), and import price inflation less

[11] Specifically, we first consider 11 different experiments, each using 20,000 draws and a modestly different value of b_{11}, based on a grid of b_{11} values centered around the asymptotically-determined value for each given R, P combination. We then pick the coefficient value that yields the lowest (in absolute value) average (across draws) difference in MSEs. We then consider a second set of 21 experiments, with a more refined grid of coefficient values and 200,000 draws. The coefficient value that yields the smallest (absolute) difference in MSEs in this second set of experiments is then used as the coefficient in the DGP simulated for the purpose of evaluating test properties.

core inflation ($x_{3,t}$).[12] Based on these data, DGP 2 takes the form

$$\begin{aligned} y_{t+1} &= -0.4y_t - 0.1y_{t-1} + b_{11}x_{1,t} + b_{12}x_{2,t} + b_{13}x_{3,t} + u_{t+1} \\ x_{1,t+1} &= 0.7x_{1,t} + v_{1,t+1} \\ x_{2,t+1} &= 0.9x_{2,t} - 0.2x_{2,t-1} + v_{2,t+1} \\ x_{3,t+1} &= 1.1x_{3,t} - 0.3x_{3,t-1} + v_{3,t+1} \end{aligned} \tag{11}$$

$$\text{var}\begin{pmatrix} u_t \\ v_{1,t+1} \\ v_{2,t+1} \\ v_{3,t+1} \end{pmatrix} = \begin{pmatrix} 0.8 & & & \\ 0.0 & 0.3 & & \\ -0.1 & 0.0 & 2.2 & \\ 0.5 & 0.1 & 0.8 & 9.0 \end{pmatrix}.$$

In DGP 2 experiments, the null (restricted) and alternative (unrestricted) forecasting models take the following forms, respectively:

$$y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + u_{0,t+1}. \tag{12}$$

$$y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 x_{1,t} + \beta_4 x_{2,t} + \beta_5 x_{3,t} + u_{1,t+1}. \tag{13}$$

As with DGP 1, we consider experiments with three different settings of the set of b_{ij} coefficients. In the Table 2 simulations of DGP 2, all of the b_{ij} coefficients are set to zero, to make the competing forecasting models equally accurate in population. In another set of experiments (Table 5), the coefficients are set at $b_{11} = 0.3$, $b_{12} = 0.1$, and $b_{13} = .015$ (roughly their empirical values). With these values, the alternative model is expected to be more accurate than the null. In the Tables 3 and 4 experiments, the values of the b_{ij} coefficients from the Table 5 experiments are multiplied by a constant less than 1, such that the null and alternative models are expected to be equally accurate, on average, over the forecast sample. As described for DGP 1, we used the asymptotic approximations developed in Clark and McCracken (2009b) and preliminary Monte Carlo experiments to determine the appropriate coefficient scaling for each R, P combination. For example, in the experiment with $R = P = 100$, the asymptotic calculations yield initial DGP 2 coefficient values of $b_{11} = 0.1327$, $b_{12} = 0.0442$, and $b_{13} = 0.0066$. The Monte Carlo refinement yields coefficient settings of $b_{11} = 0.1471$, $b_{12} = 0.0490$, and $b_{13} = 0.0074$, which are the values used to generate the results given in Tables 3 and 4 for $R = P = 100$. This Monte Carlo refinement yields models that appear to truly be equally accurate in a finite sample, with

[12] To simplify the lag structure necessary for reasonable forecasting models, the inflation rates used in forming variables $x_{2,t}$ and $x_{3,t}$ are computed as two-quarter averages.

average MSE differences (model 1 less model 2) that are at most 0.001 percentage point (the population error variance for the DGP equation is 0.8).

6.2 Results

The results presented below focus on a nominal size of 10 percent (results for 5 percent are qualitatively similar) and one-sided MSE-t tests. The tests are one-sided because, in light of the principle of parsimony familiar to forecasters, we take the small model as the null and consider only rejections of the null in favor of the alternative. However, Giacomini and White (2006) permit rejections of the larger model in favor of the smaller and conduct two-sided tests. Accordingly, we also report some results based on two-sided tests.

For each experiment, the MSE-t test is compared against three different critical values. The first, appropriate for the null hypothesis of equal accuracy in population, is computed with a *fixed regressor: population EPA* bootstrap developed in Clark and McCracken (2009b) (described here in Section 3.2). The second, appropriate for the null hypothesis of equal accuracy in the finite sample, is computed with a *fixed regressor: finite-sample EPA* developed in Clark and McCracken (2009b) (described in Section 4.2). Testing for equal accuracy in the finite sample effectively raises the bar relative to testing for equal accuracy in population (literally, yielding higher right-tail critical values from the finite-sample EPA bootstrap than from the population EPA bootstrap). Finally, the test is compared against standard normal critical values, which should be appropriate under the null of equal accuracy in a finite sample, based on the asymptotics of Giacomini and White (2006).

6.2.1 Equal accuracy at the population level

Table 2 presents Monte Carlo results for DGPs in which, in population, the competing forecasting models are equally accurate. In these DGPs, the x variables considered have no predictive content for y. As a result, while the models are equally accurate in population, in a finite sample the null forecasting model should be expected to be more accurate. In these experiments, using critical values from the population EPA bootstrap consistently yields rejection rates of about the nominal size of 10 percent. Across all of the experiments, size ranges from 9.4 to 12.6 percent, with a median of 10 percent. Other critical values typically yield rejection rates well below 10 percent, and sometimes close to 0. Under the finite-sample EPA bootstrap, rejection rates range from 0 to 7.4 percent, with a median of 2 percent. Using standard normal critical values yields rejection rates between 0 and 12.6

percent, with a median of 2 percent (in almost all cases, the test is undersized, but for a few exceptions with small P).

6.2.2 Equal accuracy in the finite sample

Table 3 presents results for DGPs in which the b_{ij} coefficients on some x variables are nonzero but small enough that, on average, the null and alternative forecasting models can be expected to be equally accurate over the sample considered. Generating critical values from the finite-sample EPA bootstrap works well: With this approach to critical values, the MSE-t test has size close to 10 percent for nearly all combinations of DGP and sample size. With DGP 1, size ranges from 9.0 to 11.7 percent, with a median of 10 percent. With DGP 2, size ranges from 9.4 to 16.6 percent; the median is 11 percent. The oversizing occurs with small R ($R = 50$) and large P/R ($P/R = 2$ and 3). More generally, with DGP 2, it is clear that, given R, the rejection rate rises with P/R. For example, with $R = 100$, the rejection rate based on the finite-sample EPA bootstrap increases from 9.8 percent at $P = 10$ to 12.9 percent at $P = 300$. (Under the recursive forecasting scheme, this pattern is much less pronounced, with the result that the finite-sample EPA bootstrap always yields nominal size of close to 10 percent.)

Using critical values from the standard normal distribution does not work as well: The MSE-t test compared against normal critical values is typically undersized, but oversized for small P. With DGP 1, size varies from 3.9 to 17.1 percent; the median rejection rate is 7 percent. With DGP 2, the size range is 5.2 to 17.7 percent, with a median of 8 percent. In contrast to the pattern noted for the finite-sample EPA bootstrap results, when the MSE-t test is compared against standard normal critical values, the rejection rate falls as P/R rises. As an example, with DGP 2 and $R = 100$, the rejection rate based on normal critical values falls from 18.8 percent at $P = 10$ to 5.8 percent at $P = 300$. This pattern runs contrary to the asymptotic results of Giacomini and White (2006), which imply that the test should be more accurate when P is large. It is possible, of course, that the asymptotics "kick in" very slowly. As a check, we ran two additional experiments with DGP 1, using (R, P) = (100, 1000) and (100, 2000), respectively. The much larger values of P are associated with very small increases in size, to 4.8 ($P = 1000$) and 5.2 percent ($P = 2000$).[13]

Testing based on the population EPA bootstrap may be seen as an unreliable indicator

[13] In the same experiments, comparing the test against critical values from the finite-sample EPA bootstrap yields rejection rates of 12.5 percent ($P = 1000$) and 13.7 percent ($P = 2000$).

of equal forecast accuracy in a finite sample. Comparing the MSE-t test against critical values from the population EPA bootstrap generally yields rejection rates far in excess of 10 percent. The rejection rate ranges from 13.9 to 50.9 percent in DGP 1 experiments and from 16.9 to 79.8 percent in DGP 2 experiments. As in prior studies such as Kilian (1999) and Clark and McCracken (2006), using a restricted VAR bootstrap, rejection rates rise as P increases. As an example, with DGP and $R = 150$, the rejection rate increases from 14.1 percent with $P = 15$ to 47.3 percent with $P = 450$.

With the performance of standard normal-based testing less favorable in our Table 3 results than the findings of Giacomini and White (2006) suggest, in Table 4 we adopt their convention and report two-sided test results. With standard normal inference, two-sided rejection rates (Table 4) are consistently higher than one-sided rejection rates (Table 3). This pattern of course suggests the finite sample distribution is shifted somewhat to the left relative to the standard normal distribution. With two-sided testing, the MSE-t test compared against normal critical values is typically undersized, but oversized for small P. However, the undersizing is less severe and the oversizing more severe than with one-sided testing. With DGP 1, two-sided size varies from 7.3 to 20.9 percent; the median rejection rate is 10 percent. With DGP 2, the size range is 6.3 to 20.8 percent, with a median of 9 percent. In contrast, using critical values from the finite-sample EPA bootstrap for a two-sided MSE-t test yields quite accurate outcomes, with size consistently close to 10 percent. Across the DGP 1 and DGP 2 experiments, the two-sided rejection rate based on the finite-sample EPA bootstrap ranges from 9.2 to 11.6 percent.

6.2.3 Alternative model most accurate

Table 5 provides results for DGPs in which the b_{ij} coefficients on some x variables are large enough that the alternative model can be expected to be more accurate than the null model — in population and in the finite sample. Not surprisingly, using critical values from the population EPA bootstrap typically yields the highest rejection rate. Across experiments, the rejection rate ranges from 18.2 to 100 percent, with a median of 72 percent. Comparing the MSE-t test against critical values estimated with the finite-sample EPA bootstrap yields somewhat lower rejection rates, ranging from 12.3 to 100 percent, with median of 40 percent. Rejection rates are systematically (for every R, P combination) lower with the finite-sample EPA bootstrap than the population EPA bootstrap because the former raises the bar, by testing the null of equal accuracy in the finite sample instead of equal accuracy in population.

Using standard normal critical values yields rejection rates between 17.1 and 100 percent, with a median of 33 percent. For small P, the normal-based tests have power comparable to the population EPA bootstrap-based tests. But for large P, the normal-based tests have relatively low power. Consider, for example, experiments with DGP 1 and $R = 100$. The rejection rate based on normal critical values rises from 21.0 percent at $P = 10$ to 67.1 percent at $P = 300$. The rejection rate based on the population EPA bootstrap is similar at $P = 10$ (22.8 percent) but much higher at $P = 300$ (97.9 percent).

7 Conclusion

This chapter provides an overview of pseudo-out-of-sample tests of unconditional predictive ability. Unconditional and conditional predictive ability differ in that the former focuses on $E(y_{T+1} - \hat{y}_{T+1})^2$, whereas the latter focuses on $E[(y_{T+1} - \hat{y}_{T+1})^2 | \Im_T]$, where $E[.|\Im_T]$ denotes the conditional expectation operator given an information set available to the forecasting agent at time T. The chapter first reviews in some detail forecast evaluation at the population level, under which estimated model coefficients converge to their population values. This analysis covers both non-nested and nested forecasting models. The chapter then reviews recent work on forecast evaluation at the finite-sample level, under which forecast accuracy is assessed on the basis of estimated model coefficients. This analysis focuses on nested models. We then provide a brief overview of other aspects of predictive ability that have been addressed in existing work or remain areas for further research. We conclude by presenting new Monte Carlo evidence on tests for equal accuracy in population versus the finite sample and on inference based on two alternative approaches developed in recent work.

References

Amisano, Gianni, and Rafaella Giacomini (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business and Economic Statistics* 25, 177-190.

Ashley, R., C.W.J. Granger and R. Schmalensee (1980), "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica*, 48, 1149-1167.

Breen, William, Lawrence R. Glosten, and Ravi Jagannathan (1989), "Economic Significance of Predictable Variations in Stock Index Returns," *The Journal of Finance* 44, 1177-1189.

Chong, Yock Y., and David F. Hendry (1986), "Econometric Evaluation of Linear Macroeconomic Models," *Review of Economic Studies* 53, 671-690.

Christoffersen, Peter F. (1998), "Evaluating Interval Forecasts," *International Economic Review* 39, 841-862.

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

Clark, Todd E., and Michael W. McCracken (2005a), "The Power of Tests of Predictive Ability in the Presence of Structural Breaks," *Journal of Econometrics* 124, 1-31.

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

Clark, Todd E., and Michael W. McCracken (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.

Clark, Todd E., and Michael W. McCracken (2009a), "Tests of Equal Predictive Ability with Real-Time Data," *Journal of Business and Economic Statistics* 27, 441-454.

Clark, Todd E., and Michael W. McCracken (2009b), "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," Working Paper 2009-05A, Federal Reserve Bank of St. Louis, October.

Corradi, Valentina, and Norman R. Swanson (2006), "Predictive Density and Conditional Confidence Interval Accuracy Tests," *Journal of Econometrics* 135, 187-228.

Corradi, Valentina, and Norman R. Swanson (2007), "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," *International Economic Review* 48, 67-109.

Corradi, Valentina, Norman R. Swanson, and Claudia Olivetti (2001), "Predictive Ability with Cointegrated Variables," *Journal of Econometrics* 104, 315-358.

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

Giacomini, Rafaella (2010), "Testing for Conditional Predictive Ability," this volume.

Giacomini, Rafaella, and Barbara Rossi (2009), "Detecting and Predicting Forecast Break-

down," Review of Economic Studies 76, 669-705.

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

Hansen, Peter R. (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics* 23, 365-80.

Hubrich, Kirstin, and Kenneth D. West (2009), "Forecast Evaluation of Small Nested Model Sets," *Journal of Applied Econometrics*, forthcoming.

Inoue, Atsushi, and Lutz Kilian (2004), "In–Sample or Out–of–Sample Tests of Predictability? Which One Should We Use?" *Econometric Reviews* 23, pp. 371-402.

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

Kuan, Chung-Ming, and Tung Liu (1995), "Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks," *Journal of Applied Econometrics* 10, 347-364.

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

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

McCracken, Michael W. (2007), "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics* 140, 719-752.

Newey, W. K., West, K. D. (1987), "A simple, positive semi–definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55:703–708.

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

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

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

West, Kenneth D. (2006), "Forecast Evaluation," in *Handbook of Economic Forecasting*, Elliott G., Granger C.W.J., Timmermann, A. (eds), North Holland.

West, Kenneth D. and Michael McCracken (1998), "Regression-based Tests of Predictive Ability," *International Economic Review* 39, 817-840.

White, Halbert (2000), "A Reality Check For Data Snooping," *Econometrica* 68, 1097-1127.

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

Table 2. Monte Carlo Rejection Rates, Equal Population-Level Accuracy

(one-sided MSE-t test, 1-step ahead rolling forecasts, nominal size = 10%)

	DGP 1			DGP 2		
	fixed regressor: population EPA	fixed regressor: finite-sample EPA	standard normal	fixed regressor: population EPA	fixed regressor: finite-sample EPA	standard normal
R=50, P=5	0.110	0.074	0.126	0.126	0.064	0.113
R=50, P=10	0.116	0.061	0.087	0.118	0.044	0.061
R=50, P=25	0.107	0.040	0.041	0.115	0.025	0.023
R=50, P=50	0.102	0.023	0.015	0.113	0.009	0.005
R=50, P=100	0.104	0.008	0.004	0.109	0.002	0.001
R=50, P=150	0.103	0.005	0.002	0.110	0.001	0.000
R=100, P=10	0.105	0.069	0.096	0.116	0.059	0.083
R=100, P=20	0.101	0.054	0.063	0.115	0.044	0.050
R=100, P=50	0.105	0.039	0.033	0.103	0.022	0.018
R=100, P=100	0.106	0.025	0.014	0.109	0.010	0.005
R=100, P=200	0.105	0.011	0.004	0.103	0.003	0.001
R=100, P=300	0.108	0.006	0.001	0.096	0.001	0.000
R=150, P=15	0.101	0.065	0.083	0.107	0.057	0.071
R=150, P=30	0.098	0.053	0.056	0.102	0.039	0.041
R=150, P=75	0.101	0.039	0.030	0.105	0.023	0.019
R=150, P=150	0.102	0.025	0.013	0.099	0.009	0.005
R=150, P=300	0.098	0.007	0.003	0.099	0.002	0.000
R=150, P=450	0.100	0.005	0.001	0.104	0.000	0.000
R=200, P=20	0.098	0.062	0.074	0.105	0.054	0.065
R=200, P=40	0.097	0.052	0.052	0.104	0.041	0.042
R=200, P=100	0.098	0.037	0.028	0.102	0.023	0.017
R=200, P=200	0.098	0.023	0.012	0.101	0.009	0.004
R=200, P=400	0.100	0.011	0.002	0.103	0.001	0.000
R=200, P=600	0.102	0.005	0.001	0.094	0.001	0.000

Notes:

1. The data-generating processes are defined in equations (8) and (11). In these experiments, the coefficients $b_{ij} = 0$ for all i, j, such that the competing forecasting models are equally accurate in population, but the null forecasting model should be expected to be more accurate in the finite-sample.
2. For each artificial data set, forecasts of y_{t+1} are formed recursively using estimates of equations (9) and (10) in the case of the DGP 1 experiments and equations (12) and (13) in the case of the DGP 2 experiments. These forecasts are then used to form the t-test for equal MSE, given in equation (3). R and P refer to the number of in–sample observations and forecasts, respectively.
3. In each Monte Carlo replication, the simulated test statistics are compared against standard normal and bootstrap critical values, using a significance level of 10%. Sections 3.2 and 4.2 describe the bootstrap procedures.
4. The number of Monte Carlo simulations is 5000; the number of bootstrap draws is 499.

Table 3. Monte Carlo Rejection Rates, Equal Finite-Sample Accuracy
(one-sided MSE-t test, 1-step ahead rolling forecasts, nominal size = 10%)

	DGP 1			DGP 2		
	fixed regressor: population EPA	fixed regressor: finite-sample EPA	standard normal	fixed regressor: population EPA	fixed regressor: finite-sample EPA	standard normal
R=50, P=5	0.148	0.101	0.171	0.197	0.104	0.177
R=50, P=10	0.171	0.097	0.132	0.234	0.102	0.134
R=50, P=25	0.215	0.102	0.100	0.346	0.115	0.104
R=50, P=50	0.284	0.101	0.072	0.493	0.126	0.082
R=50, P=100	0.405	0.104	0.053	0.674	0.143	0.070
R=50, P=150	0.509	0.117	0.051	0.798	0.166	0.068
R=100, P=10	0.141	0.093	0.129	0.182	0.098	0.138
R=100, P=20	0.157	0.092	0.106	0.230	0.105	0.119
R=100, P=50	0.205	0.095	0.080	0.320	0.107	0.095
R=100, P=100	0.287	0.100	0.063	0.459	0.108	0.071
R=100, P=200	0.403	0.103	0.046	0.649	0.127	0.061
R=100, P=300	0.490	0.112	0.046	0.761	0.129	0.058
R=150, P=15	0.141	0.098	0.119	0.178	0.099	0.123
R=150, P=30	0.158	0.090	0.094	0.211	0.095	0.100
R=150, P=75	0.216	0.100	0.081	0.305	0.099	0.082
R=150, P=150	0.271	0.092	0.055	0.448	0.112	0.070
R=150, P=300	0.398	0.105	0.051	0.632	0.111	0.052
R=150, P=450	0.473	0.097	0.039	0.753	0.121	0.053
R=200, P=20	0.139	0.094	0.109	0.169	0.098	0.113
R=200, P=40	0.156	0.091	0.091	0.204	0.094	0.098
R=200, P=100	0.203	0.092	0.073	0.316	0.105	0.085
R=200, P=200	0.287	0.101	0.057	0.437	0.101	0.063
R=200, P=400	0.384	0.094	0.040	0.630	0.107	0.053
R=200, P=600	0.471	0.099	0.040	0.749	0.117	0.052

Notes:
1. See the notes to Table 2.
2. In these experiments, the DGP coefficients $b_{ij} = 0$ are scaled such that the null and alternative models are expected to be equally accurate in the finite-sample.

Table 4. Monte Carlo Rejection Rates, Equal Finite-Sample Accuracy: Two-Sided
(two-sided MSE-t test, 1-step ahead rolling forecasts, nominal size = 10%)

	DGP 1			DGP 2		
	fixed regressor: population EPA	fixed regressor: finite-sample EPA	standard normal	fixed regressor: population EPA	fixed regressor: finite-sample EPA	standard normal
R=50, P=5	0.131	0.115	0.209	0.150	0.105	0.208
R=50, P=10	0.136	0.099	0.146	0.168	0.099	0.147
R=50, P=25	0.161	0.100	0.117	0.241	0.102	0.115
R=50, P=50	0.213	0.097	0.101	0.369	0.103	0.095
R=50, P=100	0.300	0.093	0.086	0.555	0.106	0.078
R=50, P=150	0.389	0.095	0.074	0.707	0.116	0.073
R=100, P=10	0.126	0.108	0.148	0.142	0.108	0.153
R=100, P=20	0.126	0.100	0.122	0.165	0.107	0.128
R=100, P=50	0.154	0.099	0.109	0.220	0.103	0.104
R=100, P=100	0.209	0.099	0.096	0.335	0.099	0.082
R=100, P=200	0.298	0.092	0.081	0.529	0.101	0.070
R=100, P=300	0.380	0.099	0.077	0.657	0.103	0.067
R=150, P=15	0.126	0.108	0.135	0.135	0.105	0.135
R=150, P=30	0.127	0.100	0.115	0.151	0.101	0.114
R=150, P=75	0.159	0.102	0.107	0.209	0.100	0.096
R=150, P=150	0.193	0.095	0.096	0.328	0.100	0.083
R=150, P=300	0.294	0.103	0.087	0.511	0.094	0.067
R=150, P=450	0.367	0.096	0.073	0.648	0.102	0.067
R=200, P=20	0.121	0.106	0.126	0.130	0.102	0.121
R=200, P=40	0.127	0.104	0.117	0.147	0.105	0.112
R=200, P=100	0.150	0.095	0.101	0.216	0.097	0.092
R=200, P=200	0.204	0.099	0.095	0.320	0.099	0.085
R=200, P=400	0.276	0.095	0.081	0.510	0.099	0.070
R=200, P=600	0.357	0.100	0.074	0.641	0.101	0.063

Notes:
1. See the notes to Table 2.
2. In these experiments, the DGP coefficients $b_{ij} = 0$ are scaled such that the null and alternative models are expected to be equally accurate in the finite-sample.
3. In these experiments, the MSE-t test is compared against two-sided critical values, rather than one-sided critical values as in Tables 2, 3, and 5.

Table 5. Monte Carlo Rejection Rates, Alternative Model Best
(one-sided MSE-t test, 1-step ahead rolling forecasts, nominal size = 10%)

	DGP 1			DGP 2		
	fixed regressor: population EPA	fixed regressor: finite-sample EPA	standard normal	fixed regressor: population EPA	fixed regressor: finite-sample EPA	standard normal
R=50, P=5	0.182	0.123	0.201	0.230	0.124	0.206
R=50, P=10	0.237	0.145	0.189	0.305	0.139	0.179
R=50, P=25	0.345	0.176	0.173	0.480	0.189	0.175
R=50, P=50	0.506	0.227	0.171	0.683	0.249	0.176
R=50, P=100	0.728	0.328	0.208	0.875	0.361	0.212
R=50, P=150	0.822	0.409	0.250	0.956	0.457	0.254
R=100, P=10	0.228	0.157	0.210	0.302	0.178	0.231
R=100, P=20	0.310	0.198	0.222	0.430	0.236	0.259
R=100, P=50	0.516	0.309	0.275	0.695	0.379	0.346
R=100, P=100	0.739	0.443	0.338	0.913	0.584	0.482
R=100, P=200	0.929	0.679	0.523	0.994	0.837	0.709
R=100, P=300	0.979	0.816	0.671	0.999	0.931	0.849
R=150, P=15	0.273	0.195	0.233	0.349	0.221	0.261
R=150, P=30	0.377	0.252	0.262	0.525	0.320	0.331
R=150, P=75	0.617	0.406	0.362	0.822	0.546	0.500
R=150, P=150	0.852	0.621	0.504	0.977	0.802	0.720
R=150, P=300	0.984	0.866	0.751	1.000	0.972	0.938
R=150, P=450	0.998	0.961	0.905	1.000	0.997	0.990
R=200, P=20	0.293	0.217	0.241	0.412	0.273	0.307
R=200, P=40	0.428	0.300	0.298	0.594	0.394	0.401
R=200, P=100	0.710	0.501	0.448	0.889	0.668	0.624
R=200, P=200	0.920	0.745	0.638	0.994	0.912	0.860
R=200, P=400	0.997	0.952	0.893	1.000	0.997	0.990
R=200, P=600	1.000	0.993	0.972	1.000	1.000	1.000

Notes:
1. See the notes to Table 2.
2. In these experiments, the coefficients $b_{ij} = 0$ are set to values (given in section 6.1) large enough that the alternative model is expected to be more accurate than the null model.