Evaluating the Predictability of Exchange Rates using Long Horizon Regressions: Mind Your p's and q's! *

Michael W. McCracken
Assistant Professor of Economics
University of Missouri-Columbia

Stephen Sapp
Assistant Professor of Finance
University of Western Ontario

Version: June 11, 2004

Abstract

Since the breakdown of the Bretton Woods agreement, researchers have used a wide variety of structural models to try to predict exchange rate movements. Unfortunately, finding consistent evidence that these models outperform a random walk has proven elusive. In this paper we investigate the impact different methods of inference may have had on these conclusions. Using p-values based on recently developed tests of forecast accuracy and encompassing, as well as q-values designed to mitigate multiple testing problems, we provide stronger evidence consistent with these models having superior predictive ability. Our results suggest that previous studies' inability to detect predictive ability may have been influenced by the statistics used and the manner in which they were employed.

Keywords: forecast evaluation, exchange rates, long-horizon regression.

J.E.L. Nos.: C52, C53, F31

McCracken: Dept. of Economics, University of Missouri-Columbia, 118 Professional Building, Columbia, MO 65211, USA, mccrackenm@missouri.edu. *Sapp(corresponding author)*: University of Western Ontario, Ivey School of Business, 1501 Richmond Street N, London, ON N6A 3K7, Canada, ssapp@ivey.uwo.ca.

Evaluating the Predictability of Exchange Rates using Long Horizon Regressions: Mind Your p's and q's!

Michael W. McCracken
Assistant Professor of Economics
University of Missouri-Columbia

Stephen Sapp
Assistant Professor of Finance
University of Western Ontario

Abstract

Since the breakdown of the Bretton Woods agreement, researchers have used a wide variety of structural models to try to predict exchange rate movements. Unfortunately, finding consistent evidence that these models outperform a random walk has proven elusive. In this paper we investigate the impact different methods of inference may have had on these conclusions. Using p-values based on recently developed tests of forecast accuracy and encompassing, as well as q-values designed to mitigate multiple testing problems, we provide stronger evidence consistent with these models having superior predictive ability. Our results suggest that previous studies' inability to detect predictive ability may have been influenced by the statistics used and the manner in which they were employed.

1. Introduction

When the Bretton Woods Agreement broke down in 1973, most of the large industrialized countries allowed their exchange rates to float against one another. Because this was the first widespread floating of exchange rates in over fifty years, researchers were motivated to develop and estimate empirical models to understand the observed movements in exchange rates. Although preliminary studies had some success at explaining exchange rates, by the early 1980's many of the early successes were being overturned.

One of the most significant negative results was Meese and Rogoff (1983a, b). They analyze the predictive ability of a series of linear structural exchange rate models and found that none was able to consistently outperform a simple random walk across various exchange rates and forecast horizons. Despite the robustness of this result (e.g. Mark and Sul (2002), Rapach and Wohar (2002) and Faust, Rogers and Wright (2003)), there is some evidence of linear structural models outperforming random walk models (e.g. Chinn and Meese (1995), Mark (1995), and MacDonald and Marsh (1997)). Recent work using non-linear models has also shown promise (e.g. Taylor, Peel and Sarno (2001), Cheung, Chinn and Pascual (2002), Clarida, Sarno, Taylor and Valente (2003), and Kilian and Taylor (2003)).

Because the original results of Meese and Rogoff (1983a, b) have yet to be convincingly overturned, we investigate the role that the method of inference may have played in determining whether or not structural exchange rate models exhibit superior predictive ability to the random walk model. Much of the existing literature, including Meese and Rogoff (1988) and recent extensions such as Cheung, Chinn and Pascual (2002), implement a t-type test of equal forecast accuracy recently associated with Diebold and Mariano (1995). Inference is conducted treating these statistics as asymptotically standard normal. As discussed in Section 2, such an approximation is valid when comparing the forecast accuracy of two non-nested models but is

invalid when comparing two nested models. Since the structural models typically nest the random walk model, using normal critical values is inappropriate and the resulting p-values do not accurately reflect the significance of the test statistic.

In this paper we evaluate the predictive ability of linear structural exchange rate models, including the monetary model (Frenkel (1976), Mussa (1976) and Bilson (1978)), relative to the random walk model with drift using test statistics explicitly designed for an out-of-sample comparison of nested models. Building upon the results in West (1996), McCracken (2000) and Clark and McCracken (2001) derive the limiting distributions of four out-of-sample tests of forecast accuracy and encompassing for one-step ahead forecasts from nested models. Clark and McCracken (2003a) extend their results to allow multi-step forecasts from long-horizon regressions. In related work, Chao, Corradi and Swanson (2001) derive a test of forecast encompassing that is applicable when one-step ahead forecasts are constructed from either nested or non-nested models. In Section 2 of this paper, we provide an extension of their test that allows for forecasts from longer horizons.

Each of the encompassing tests associated with Chao, Corradi and Swanson (2001) are asymptotically chi-square and hence asymptotically valid p-values are readily constructed using the relevant tables. Since the remaining tests have nonstandard limiting distributions that are usually dependent upon unknown nuisance parameters, we follow Clark and McCracken (2003a) in using a bootstrap similar to that in Kilian (1999) to estimate asymptotically valid critical values and construct asymptotically valid p-values.

Another reason for using these new tests is that Clark and McCracken (2001, 2003a, b) provide analytical, Monte Carlo and empirical evidence that some out-of-sample tests of predictive ability have greater power than others. In particular they show that the commonly used t-type tests of either forecast accuracy or encompassing have lower power than their F-type

counterparts. Since much of the literature in this area, including Meese and Rogoff (1988), Mark (1995), Kilian (1999) and Cheung, Chinn and Pascual (2002) focus on t-type tests of forecast accuracy, it may be that their results are due in part to using tests that have low power.[1]

Using these new tests we begin our analysis as in Mark (1995) by investigating the predictive ability of the monetary model for quarterly US dollar exchange rates with the German mark, Canadian dollar, Japanese yen and Swiss franc. Using data from 1973 to 1991, Mark (1995) finds evidence of predictive ability for the monetary model. He also suggests that the predictive ability should increase as the forecast horizon increases and the results would be even stronger in a longer data series. Due to the striking nature of his results, this study has been the focus of much subsequent work. For example, Kilian (1999) replicates his analysis using a slightly modified technique and longer time series but comes to a different conclusion. Faust, Rogers and Wright (2003) suggest that his results were strongly influenced by a fortuitous choice of time period. Berben and van Dijk (1998) and Berkowitz and Giorgianni (2001) found a potentially large impact of his assumption of cointegration on both the estimation technique and the distribution of the test statistics.[2] Rapach and Wohar (2002) consider a century of annual data rather than post-Bretton Woods quarterly data.

Based on quarterly, post-Bretton Woods data from 1973 to 1998 we provide evidence that many structural models exhibit more predictability than previously believed. In our tests we use the random walk with drift as our benchmark to more clearly isolate the marginal contribution of the fundamental factors as suggested in Kilian (1999). We find that for the monetary model, the new F-type tests of equal forecast accuracy indicate more short-horizon predictability (especially for Germany) and more long-horizon predictability (especially for Canada and Switzerland) than is found using the t-type test of forecast accuracy used in Mark (1995) and Kilian (1999). The results are robust across the linear structural exchange rate models considered. The evidence of

predictive ability is reinforced when we consider the tests of forecast encompassing. Again we find that the F-type tests provide stronger evidence of predictive ability than the t-type tests though the evidence is less uniform than that for the tests of forecast accuracy.

This evidence of predictive ability is subject to the criticism that we have a multiple testing problem. This issue arises because we are conducting inference on 4 distinct models of 4 bilateral exchange rates at 5 horizons using 5 test statistics yielding 400 separate test statistics. In that light it is not surprising that we would find predictive ability using p-value thresholds of 5% and 10%. However, we do not reach our conclusions based solely on the use of p-value thresholds. We strengthen our arguments using procedures recently receiving attention in the statistical genetics literature where literally thousands of genes are often tested for specific properties. These procedures involve constructing not only asymptotically valid estimates of p-values but also *q-values*. Both p- and q-values can be interpreted as measures of a statistic's significance, each from a different perspective. For example, if a test statistic has a p-value of 5% we would expect that among a random sample of pairs of hypotheses and statistics from the same population as the statistic, on average 5% of those hypotheses that are null will have statistics that reject. Conversely, if a statistic has a q-value of 5% we expect that on average 5% of the statistics that reject actually correspond to null hypotheses. We construct these q-values using methods discussed in Storey (2002). A more complete description of q-values is provided in Section 2.

The remainder develops as follows. In Section two we discuss the tests used to detect predictive ability as well as the assumptions necessary for their application. Section three discusses the structural exchange rate models we use for forecasting and other details about our data and empirical methods. Section four presents empirical evidence on the predictive ability of the structural exchange rate models. The final section concludes.

2. Testing Procedures and Environment

The framework we use to evaluate predictive ability is similar in spirit to that used in Meese and Rogoff (1983a, b) but is closest to that in Mark (1995) and Kilian (1999). That is, we consider whether structural model-based long-horizon regressions of log-exchange rates provide superior predictive ability to a random walk with drift in log-exchange rates using out-of-sample methods. The intuition for using out-of-sample methods is that if one of the models can be shown to predict better than another in such a forecasting exercise then it may continue to do so as more data become available.

Environment

The sample of covariance stationary observables $\{y_t, x_{2,t}^{'}\}_{t=1}^{T}$ includes a scalar random variable y_t to be predicted (e.g. the changes in log exchange rates) and a ($k_1 + k_2 = k \times 1$) vector of potential predictors $x_{2,t} = (x_{1,t}^{'}, x_{22,t}^{'})^{'}$. The sample is divided into in-sample and out-of-sample portions. The in-sample portion spans observations 1 to R. Letting $P - \tau + 1$ denote the number of τ-step ($1 \leq \tau$) ahead forecasts, the out-of-sample observations span $R + \tau$ through $R + P$. The total number of observations in the sample is $R + P = T$.

Forecasts of $y_{t+\tau}$, $t = R,\ldots,T - \tau$, are generated using two linear models of the form $x_{i,t}^{'}\beta_i^{*}$, $i = 1, 2$. The sequence of parameter estimates used to construct the forecasts are estimated recursively by OLS. We denote the estimated τ-step ahead forecast errors as $\hat{u}_{1,t+\tau} = y_{t+\tau} - x_{1,t}^{'}\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+\tau} = y_{t+\tau} - x_{2,t}^{'}\hat{\beta}_{2,t}$. These functions, when evaluated at the population parameters, are denoted $u_{1,t+\tau} = y_{t+\tau} - x_{1,t}^{'}\beta_1^{*}$ and $u_{2,t+\tau} = y_{t+\tau} - x_{2,t}^{'}\beta_2^{*}$. Note that in our notation, under the null hypothesis of either equal forecast accuracy or forecast encompassing, model 1 is nested within

model 2 and hence $u_{1,t+\tau} = u_{2,t+\tau}$.

In each of the models, the dependent variable $y_{t+\tau}$ is the τ-step difference in the log-exchange rate series s_t. Hence $y_{t+\tau} = s_{t+\tau} - s_t$. To clarify our definitions of $x_{1,t}$ and $x_{2,t}$ suppose we are comparing the forecast accuracy of the random walk with drift model of exchange rates with the monetary model as considered in Kilian (1999). In this case $x_{1,t} = 1$ and $x_{2,t} = (1, z_t)'$ for $z_t = s_t - (m_t - m_t^*) + (y_t - y_t^*)$ where $m_t - m_t^*$ denotes the logarithm of the ratio of the US money supply to the foreign money supply and $y_t - y_t^*$ is the logarithm of the ratio of US to foreign real income.[3]

Test Statistics

Because these test statistics are discussed in detail elsewhere, here we only present a brief overview. The first two statistics are used to test for equal MSE. Let $\hat{d}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}$, $\bar{d} = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau} \hat{d}_{t+\tau} = MSE_1 - MSE_2$, $\hat{\Gamma}_{dd}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T-\tau} \hat{d}_{t+\tau}\hat{d}_{t+\tau-j}$ for $j \geq 0$ and $\hat{\Gamma}_{dd}(j) = \hat{\Gamma}_{dd}(-j)$. If we estimate the long-run covariance of $d_{t+\tau}$ using a kernel-based estimator with kernel function K(.), bandwidth parameter M and maximum number of lags $\bar{j}$ so that $\hat{S}_{dd} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\hat{\Gamma}_{dd}(j)$ the test statistics take the form

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

$$MSE\text{-}F = (P-\tau+1) \frac{\bar{d}}{MSE_2} \qquad (2)$$

Under the null that the mean square error associated with model 1 is the same as that for model 2, the expected difference between $u^2_{1,t+\tau}$ and $u^2_{2,t+\tau}$ is zero. Under the alternative the mean square error associated with model 2 will be smaller than that for model 1. Hence when constructing tests of equal forecast accuracy using either of these statistics we use critical values

chosen from the upper tail of the null limiting distribution. Since the limiting distributions of these two statistics are non-standard and depend upon unknown nuisance parameters, we rely on the results in Clark and McCracken (2003a) to motivate our use of a bootstrap similar to that in Kilian (1999) to estimate asymptotically valid critical values and corresponding p-values.

The remaining three statistics are used to test for forecast encompassing. The first two of these are motivated by a t-type statistic used by Harvey, Leybourne, and Newbold (1998) to test for forecast encompassing between two non-nested models. Let $\hat{c}_{t+\tau} = \hat{u}_{1,t+\tau}(\hat{u}_{1,t+\tau}-\hat{u}_{2,t+\tau})$, $\bar{c} = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\hat{c}_{t+\tau}$, $\hat{\Gamma}_{cc}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T-\tau}\hat{c}_{t+\tau}\hat{c}_{t+\tau-j}$ for $j \geq 0$ and $\hat{\Gamma}_{cc}(j) = \hat{\Gamma}_{cc}(-j)$. If we estimate the long-run covariance of $c_{t+\tau}$ using a kernel-based estimator with kernel function K(.), bandwidth parameter M and maximum number of lags $\bar{j}$ so that $\hat{S}_{cc} = \sum_{j=-\bar{j}}^{\bar{j}}K(j/M)\hat{\Gamma}_{cc}(j)$ the test statistics take the form

$$\text{ENC-t} = (P-\tau+1)^{1/2}\frac{\bar{c}}{\sqrt{\hat{S}_{cc}}} \tag{3}$$

$$\text{ENC-F} = (P-\tau+1)\frac{\bar{c}}{MSE_2} \tag{4}$$

Under the null that the forecast from model 1 encompasses that of model 2, the covariance between $u_{1,t+\tau}$ and $u_{1,t+\tau} - u_{2,t+\tau}$ will be less than or equal to zero. Under the alternative that model 2 contains added information, the covariance should be positive. Hence when constructing tests of forecast encompassing using either of these statistics we use critical values chosen from the upper tail of the null limiting distribution. Since the limiting distributions of these two statistics are non-standard and depend upon unknown nuisance parameters, we continue to use a bootstrap procedure to estimate asymptotically valid critical values and corresponding p-values.

The final encompassing test we consider was developed by Chao, Corradi and Swanson (2001). This statistic is closely related to one used by Chong and Hendry (1986) to test for forecast encompassing between two non-nested models. Let $\hat{h}_{t+\tau} = \hat{u}_{1,t+\tau}x_{1,t}$, $\hat{b}_{t+\tau} = \hat{u}_{1,t+\tau}x_{22,t}$, $\bar{b} = (P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}\hat{b}_{t+\tau}$, $\hat{F} = -(P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}x_{22,t}x_{1,t}'$ and $\hat{B} = ((P-\tau+1)^{-1}\sum_{t=R}^{T-\tau}x_{1,t}x_{1,t}')^{-1}$. Let $\hat{\Gamma}_{bb}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T-\tau}\hat{b}_{t+\tau}\hat{b}_{t+\tau-j}$, $\hat{\Gamma}_{hh}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T-\tau}\hat{h}_{t+\tau}\hat{h}'_{t+\tau-j}$ and $\hat{\Gamma}_{bh}(j) = (P-\tau+1)^{-1}\sum_{t=R+j}^{T-\tau}\hat{b}_{t+\tau}\hat{h}'_{t+\tau-j}$ for $j \geq 0$ with $\hat{\Gamma}_{bb}(j) = \hat{\Gamma}_{bb}(-j)$, $\hat{\Gamma}_{hh}(j) = \hat{\Gamma}_{hh}(-j)$ and $\hat{\Gamma}_{bh}(j) = \hat{\Gamma}_{bh}(-j)$. If we estimate the long-run covariance of $(b_{t+\tau}', h_{t+\tau}')'$ using a kernel-based estimator with kernel function K(.), bandwidth parameter M and maximum number of lags $\bar{j}$ so that $\hat{S}_{bb} = \sum_{j=-\bar{j}}^{\bar{j}}K(j/M)\hat{\Gamma}_{bb}(j)$, $\hat{S}_{hh} = \sum_{j=-\bar{j}}^{\bar{j}}K(j/M)\hat{\Gamma}_{hh}(j)$ and $\hat{S}_{bh} = \sum_{j=-\bar{j}}^{\bar{j}}K(j/M)\hat{\Gamma}_{bh}(j)$ the test statistic takes the form

$$CCS = (P-\tau+1)\bar{b}'\hat{\Omega}^{-1}\bar{b} \qquad (5)$$

where $\hat{\Omega} = \hat{S}_{bb} + \hat{\lambda}_{bh}(\hat{F}\hat{B}\hat{S}_{bh}' + \hat{S}_{bh}\hat{B}'\hat{F}') + \hat{\lambda}_{bb}\hat{F}\hat{B}\hat{S}_{hh}\hat{B}'\hat{F}'$, $\hat{\pi} = (P-\tau+1)/R$, $\hat{\lambda}_{bh} = 1 - \hat{\pi}^{-1}\ln(1+\hat{\pi})$ and $\hat{\lambda}_{hh} = 2[1 - \hat{\pi}^{-1}\ln(1+\hat{\pi})]$. Under the null that forecasts from model 1 encompass those of model 2, the covariance between $u_{1,t+\tau}$ and $x_{22,t}$ will be zero. Under the alternative that model 2 contains added information, the covariance should be non-zero. Since the null limiting distribution of the statistic is $\chi^2(k_2)$ we choose critical values and construct p-values using the appropriate tables.

Inference

Since we are using multiple test statistics for each exchange rate, forecast horizon and structural model it is not surprising that in Section four we find numerous instances in which the p-values are less than 10%. To improve the reliability of our inference for each of the tests, we do not reject a particular null hypothesis simply because it's associated p-value happens to be

below a pre-chosen threshold, α, such as 10%. Instead we make a decision to reject the null for a particular test based upon its' corresponding p- and q-value respectively. Although the use of q-values is increasingly common in the statistics literature, and in particular those applications related to genetics where literally thousands of genes are being tested for some feature (e.g. Storey and Tibshirani (2003)), q-values are relatively unknown in the economics literature and hence we provide a brief description below.[4]

Consider an experiment in which m distinct tests are being conducted. Among these suppose that m_0 and m_1 ($m = m_0 + m_1$) denote the number of instances in which the null and alternative are true respectively. If we let S denote the number of rejections and F and T denote the number of false and true rejections we obtain the following table.

	Reject	Fail to reject	Total
Null True	F	$m_0 - F$	m_0
Alternative True	T	$m_1 - T$	m_1
Total	S	$m - S$	m

In the standard situation where only a single test is being performed ($m = 1$) one selects a rejection rule that maximizes the power of the test, $E(T/m_1) = Pr(T = 1)$, when $m_1 = 1$ but controls the false positive rate or probability of a type I error, $E(F/m_0) = Pr(F = 1)$, below some threshold α when $m_0 = 1$. In this environment, we reject the null hypothesis when the p-value associated with a test statistic is less than or equal to α and fail to reject otherwise.

In a multiple testing environment (i.e. $m \geq 2$) it is less clear how one should conduct inference and, in particular, decide which of the m hypotheses correspond to the null or alternative hypotheses. One naïve method consists of simply applying the approach designed for the single test case to each of the individual hypotheses without any regard to the existence of the other tests. However, if we do so we no longer ensure that the false positive rate $E(F/m_0)$ is less

than α but instead only ensure that $E(F/m) \leq \alpha$. This bound is extremely forgiving and can lead to too many false rejections of the null.

The Bonferroni correction gets around this problem by changing the rejection rule. When using the Bonferroni correction we reject the null for any particular test only if the corresponding p-value is less than or equal to α/m. This ensures that the false positive rate remains below our threshold, $E(F/m_0) \leq \alpha$, but does so at the expense of power. Despite the loss in power for a particular test, in applications where m_0 is large relative to m_1 or when a false rejection of the null is costly, the Bonferroni correction may be useful.

However, in applications like ours where, a priori, economic theory (Frenkel (1976), Mussa (1976) and Bilson (1978)) and the power of the test statistics (Clark and McCracken (2003b)) leads us to expect $T \geq 1$, controlling the false discovery rate, $E(F/S)$, rather than the false positive rate, $E(F/m_0)$, may be a more appealing intermediate alternative to either the liberal naïve approach or conservative Bonferroni correction. This approach, proposed by Benjamini and Hochberg (1995), is designed for situations where we are more interested in ensuring that our rejections are legitimate (e.g. F/S is small) than in guarding against one or more false positives (e.g. F/m_0 is small). As a result their method is likely to have better power than the Bonferroni correction and is likely to have fewer false positives than the naïve approach.

Building upon the concept of controlling the false discovery rate, Storey (2003) defines a test statistic, the q-value, as the minimum possible false discovery rate for which we reject the null just as we define a test's p-value as the minimum possible false positive rate for which we will reject the null. Both of these values, along with a given threshold for determining significance, permit us to describe the confidence we have in our statistical inference in several dimensions using our 2×2 table above. For example, when a p-value is less than (greater than) a prespecified

level α we categorize it as contributing to the first (second) column and hence is an element of S (m – S). In contrast, when a q-value is less than (greater than) a prespecified level α we categorize it as contributing to the second (first) row and hence is an element of m_1 (m_0).

Methods for constructing q-values and their properties are discussed in Storey (2002) as well as Storey, Taylor and Siegmund (2003). Intuitively the algorithm for calculating the q-value compares the distribution of the observed p-values from the series of tests to what one would have expected if the null were true in all cases. This requires calculating the percentage of cases we would expect to be consistent with the null, $\pi_0(\alpha) = m_0/m$, for each level of significance, α. As the level of significance decreases (e.g. as α goes from 0.01 to 0.99), the only cases for which we can not reject the null will be when the null is, in fact, true. Consequently, our estimate for $\pi_0(1)$ is the most conservative estimate we have for m_0/m. The q-value uses this conservative estimate to determine the probability with which we will falsely reject the null hypothesis for a given p-value. Formally the q-values can be constructed using the following algorithm:

1. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ be the ordered p-values from the m tests of interest.
2. For $\lambda \in [0.01, 0.99]$ estimate $\pi_0 = m_0/m$ using $\hat{\pi}_0(\lambda) = \dfrac{\#(p_j > \lambda)}{m(1-\lambda)}$.
3. Fit a cubic spline $f^*(.)$ of $\hat{\pi}_0(\lambda)$ on λ.
4. Set $\hat{\pi}_0 = f^*(1)$.
5. Calculate $\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \dfrac{\hat{\pi}_0 m \cdot t}{\#(p_j \leq t)} = \hat{\pi}_0 \cdot p_{(m)}$.
6. For i = m-1, m-2,…, 1 calculate $\hat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \dfrac{\hat{\pi}_0 m \cdot t}{\#(p_j \leq t)} = \min\left(\dfrac{\hat{\pi}_0 m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)}) \right)$.
7. The estimated q-value for the i^{th} most significant test is $\hat{q}(p_{(i)})$.

In our study we provide both the p- and q-values for each of the 400 test statistics we construct. Construction of the asymptotically valid p-values is discussed below. Given those p-

values we implement the publicly available software QVALUE to construct the asymptotically valid q-values associated with each of the 400 statistics. A detailed discussion of the software can be found at http://faculty.washington.edu/~jstorey/qvalue/.

Before proceeding to our empirical work we should be clear about the assumptions that we are relying upon to construct the q-values. The primary assumptions are (1) the p-values are asymptotically valid, (2) the p-values satisfy certain weak dependence conditions and (3) m is large. That the p-values are asymptotically valid follows from the fact that that we use a bootstrap to account for the non-standard asymptotic distributions implied by the nested model comparisons. That we satisfy the relevant dependence conditions or that m is 'large' is less clear. Nevertheless it is reasonably well-established that smaller sample sizes and stronger dependence imply increasingly *conservative* (i.e. larger) estimates of the q-values. Formal discussions can be found in Storey, Taylor and Siegmund (2003), especially Theorem 7 and the corresponding numerical simulations. Furthermore, studies such as Storey and Tibshirani (2001) and Storey (2002) demonstrate the robustness of the q-value to sample sizes (m) varying from 500 to 10,000 using different degrees of dependence across the test statistics.

3. Empirical Methods

To facilitate the comparison between our study and the existing literature, we use models, data and forecasting procedures similar to those used in some of the most widely cited studies. By doing so we also aim to isolate the contribution that the new test statistics can make in detecting predictive ability.

As is standard in much of the literature, the models that we consider define the current value of foreign exchange in terms of fundamental macroeconomic factors.[5] The individual models we consider are all special cases of the following model

$$s_{t+\tau} - s_t = \beta_0^* + \beta_1^* z_t + u_{t+\tau} \quad (6)$$

$$z_t = s_t - f_t$$

$$f_t = \alpha_1(m_t - m_t^*) + \alpha_2(y_t - y_t^*) + \alpha_3(r_t - r_t^*) + \alpha_4(\pi_t - \pi_t^*) + \alpha_5 TB_t + \alpha_6 TB_t^*$$

where s_t is the logarithm of the current US dollar – foreign currency exchange rate, $(m_t - m_t^*)$ is the difference in the logarithms of the US money supply and the foreign money supply, $(y_t - y_t^*)$ is the difference in the logarithms of the US and foreign real income, $(r_t - r_t^*)$ is the short-term interest rate differential and $(\pi_t - \pi_t^*)$ is the expected long-run inflation differential while TB_t and TB_t^* represent the net US and foreign trade balances respectively.

Our models are largely drawn from those used in Meese and Rogoff (1983a, b). The specific models we consider are the monetary model, the flexible price monetary model (Frenkel-Bilson), the sticky-price monetary model (Dornbusch-Frankel) and the sticky-price asset model (Hooper-Morton). Each of these models can be represented using (6) by making different assumptions on the coefficients. For the monetary model (model 1), we follow Mark (1995) and Kilian (1999) by setting $\alpha_1 = 1$ and $\alpha_2 = -1$. For the other models we use commonly assumed values[6]: for the flexible-price monetary model (model 2) $\alpha_1 = 1$, $\alpha_2 = -1$, and $\alpha_3 = -1$, for the sticky-price monetary model (model 3) $\alpha_1 = 0$, $\alpha_2 = -1$, $\alpha_3 = -1$, and $\alpha_4 = 1$ and for the sticky-price asset model (model 4) $\alpha_1 = 1$, $\alpha_2 = -1$, $\alpha_3 = -0.5$, $\alpha_4 = 3$, $\alpha_5 = 0.001$ and $\alpha_6 = -0.001$.[7]

The data was chosen to conform to the structural models and to be consistent with the data used in previous studies. For the monetary model we start with the data used in Kilian (1999). It was constructed using the OECD Main Economic Indicators available from DataStream for the period from 1973:Q1 to 1997:Q4. It includes the US dollar exchange rates of the Canadian dollar, the German mark, the Japanese yen and the Swiss franc as well as the necessary fundamentals. These include the GNP and the money supply (M1) for each country. For

Switzerland we use GDP rather than GNP and for Canada we use M3 rather than M1 for the money supply due to data availability. These data are supplemented to include the data required for the estimation of the other structural models. These include the Treasury bill rates (the short-term interest rates in each country), the CPI for the inflation and trade balances in each country all from DataStream. We deseasonalize this data in a manner similar to Mark (1995) by using a rolling aggregation of the data over the past year.

With the fundamentals defined above, for any fixed model i = 1, 2, 3, 4 and horizon τ = 1, 2, 4, 8, 12 we construct our forecasts using the linear regression models

$$s_{t+\tau} - s_t = \beta^*_{0,2} + \beta^*_{1,2} z_t + u_{2,t+\tau} = x'_{2,t}\beta^*_2 + u_{2,t+\tau} \quad (7)$$

$$s_{t+\tau} - s_t = \beta^*_{0,1} + u_{1,t+\tau} = x'_{1,t}\beta^*_1 + u_{1,t+\tau} \quad (8)$$

where $x_{1,t} = 1$ and $x_{2,t} = (1, z_t)'$.[8] To evaluate predictive ability we estimate (7) and (8) by OLS $P-\tau+1$ times once each using observations $j = 1,\ldots,t$ for $t = R,\ldots,T-\tau$. For each t we then construct two τ-step ahead forecasts, $x'_{1,t}\hat{\beta}_{1,t}$ and $x'_{2,t}\hat{\beta}_{2,t}$, with corresponding forecast errors $\hat{u}_{1,t+\tau} = y_{t+\tau} - x'_{1,t}\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+\tau} = y_{t+\tau} - x'_{2,t}\hat{\beta}_{2,t}$. Given the subsequent two sequences of forecast errors we construct each of the test statistics in equations (1) – (5).

For each model, horizon and exchange rate we follow the detailed steps given below.

1) Transform the data as in Mark (1995) and Kilian (1999) for each of the models. The resulting values are multiplied by 400 to convert the relevant series to annualized percentage points.

2) Models (7) –(8) are initially estimated over the in-sample period 1973:Q1 – 1989:Q4.[9]

3) Models (7) – (8) are estimated recursively over the out-of-sample period 1990:Q1 – 1997:Q4.

4) The test statistics are calculated using (1) – (5). The long-run covariances in (1), (3) and (5) are calculated using the Newey and West (1987) estimator with a lag length of (the integer component of) 1.5τ.[10]

5) The p-values for the CCS statistic are calculated using the upper tail of the chi-square(1) distribution. The p-values associated with the MSE-t, MSE-F, ENC-t and ENC-F statistics are estimated using a bootstrap similar to that discussed in Kilian (1999). The algorithm consists of:

a) Estimate the DGP for the exchange rate and the expected exchange rates as:

$$s_t - s_{t-1} = d + \upsilon_t$$

$$f_t - f_{t-1} = a + bz_{t-1} + \sum_{j=1}^{p-1} \zeta_j \Delta s_{t-j} + \sum_{j=1}^{p-1} \gamma_j \Delta f_{t-j} + \varepsilon_t.$$

The number of lags, p, is determined using AIC.

b) Take the resulting residuals and sample them with replacement to obtain a set of bootstrap residuals, υ_t^* and ε_t^*.

c) Create a bootstrap series of the changes in log exchange rates, s_t, and changes in market fundamentals, f_t, recursively using these residuals and the estimated coefficients. To initialize the process set $z_t^* = 0$ and $\Delta s_{t-j}^* = \Delta f_{t-j}^* = 0$ for j = p-1,…,1 and discard the first 500 transients. Based on these values we calculate the bootstrap series s_t^* and f_t^* the same length as the original data series.

d) Repeat steps 2 – 4 for the bootstrap sample.

e) Repeat steps a) – d) until we have the required number of bootstrapped test statistics.

f) Use the bootstrapped test statistics to obtain an empirical distribution to calculate the p-values for the test statistics from the true data.

6) Given the p-values from step 5), we estimate the corresponding q-values using the program QVALUE discussed in Section two.

4. Empirical Results

Before discussing our results using the new test statistics, we briefly summarize the results for the monetary model from Kilian (1999). Kilian (1999) repeated the analysis of Mark (1995) and found that increasing the forecasting horizon and sample size did not significantly increase the predictive power of the models as originally hypothesized by Mark (1995). In fact, using tests based on the MSE-t statistic Kilian (1999) found only minor evidence of exchange rate predictability over the period from 1973 to 1997. Specifically, for Canada and Switzerland he found some evidence of predictive ability at the shorter forecasting horizons. However, for Germany and Japan the evidence of predictive ability was weaker. The only significant predictive power was for the Japanese yen for one period ahead (one quarter). To determine the value of our test statistics, we repeat the analysis performed by Kilian (1999) extending it to include our expanded set of models and our diverse set of test statistics.

Baseline Results

Table 1 provides the root mean square errors (RMSE) associated with the random walk with drift model as well as the ratios ($RMSE_{\text{Random Walk}}/RMSE_{\text{Structural Model}}$).[11] We immediately see that of the 80 RMSE ratios for the different structural models all but 13 of them are greater than 1 though the indicated gains from using the structural models is frequently not large. For example, the sticky-price monetary model (model 2) applied to Switzerland has ratios as low as 1.01 at the 12-quarter horizon and as high as 1.34 at the 8-quarter horizon. Even so, there are examples where the ratio is quite large and this is particularly acute for Canada. For example, the monetary model (model 1) applied to Canada has ratios that rise from 1.01 at the 1-quarter horizon to as large as 3.57 at the 12-quarter horizon. Of the 13 instances in which the ratio is less than or equal to 1, all occur at the longest two horizons. In fact all of the ratios associated

with Germany at forecast horizons of 8 and 12 quarters and all of the ratios for Japan at the 12-quarter forecast horizon are less than 1. Note that since for any fixed horizon and exchange rate, these ratios are consistently greater (or consistently less) than one nearly all of our structural models are capturing similar movements in exchange rates. For example, at the 2-quarter horizon the ratios for Japan take the values 1.07, 1.08, 1.08 and 1.07 across models 1 – 4.

The RMSE ratios provide some insight into what we should expect in the subsequent analysis. In general the ratios indicate that, with the exception of Canada, the structural models have slightly more predictive power than the random walk with drift at shorter horizons but not at longer horizons. For Canada the ratios suggests the structural models have significant predictive power at long horizons. These results are broadly consistent with the findings of Kilian (1999) but suggest that more powerful tests of forecast accuracy may be able to detect predictive ability for our structural models, especially at short horizons.

In Tables 2 – 6 we report the p- and q-values for the five out-of-sample tests of predictive ability for each model, currency and forecast horizons. The tables are for the MSE-t, MSE-F, ENC-t, ENC-F and CCS statistics respectively. Before discussing each of these in detail it is instructive to consider Figure 1. Here we provide a histogram, consisting of 40 equally spaced bins, of all 400 of the p-values from the corresponding tests. The plot is interesting because it gives a feel for whether or not a portion of the tests indicate that the structural models have predictive ability beyond that of the random walk with drift model. If the null hypothesis of no predictive ability was satisfied for all 400 tests we would expect the distribution of p-values to be approximately uniform with roughly 10 p-values in each bin. If, on the other hand, a proportion of the hypotheses were alternative we would expect a distribution that was a mixture of a uniform and a spike near zero.

The plot in Figure 1 is distinctly not uniform. Intuitively this can be seen by the fact that there are far too few p-values greater than 50% and far too many p-values less than 20%. Despite the intuitive appeal of the plot, it is far more difficult to prescribe a well defined and well behaved decision rule for determining which p-values correspond to alternative hypotheses. Figure 2 provides insight into some of the characteristics of the q-values and how they relate to the p-values. The first figure demonstrates how the number of tests demonstrating significant predictive ability increases as we allow more false positives (i.e. we allow a higher q-value). Notice that for estimated q-values greater than 0.04 the number of significant tests increases dramatically. However, the second figure suggests that the number of false positives among this group remains low.

Before discussing the individual Tables it is useful to make a few general observations regarding the p- and q-values and how they can be used to interpret the results. Building on the notation we introduced earlier, of the m = 400 tests we have 154 cases where the p-values are less than 10% ($S_{10\%}$ = 154) while 94 have p-values less than 5% ($S_{5\%}$ = 94). Similarly, of the 400 tests 338 have q-values less than 10% while 210 have q-values less than 5%. Each of the 154 (94) statistics with p-values less than 10% (5%) have q-values less than 4.9% (3.9%). In other words, using a 10% (5%) p-value threshold for rejection implies at most a 4.9% (3.9%) q-value threshold for rejection, or equivalently that at most 4.9% (3.9%) of the rejections are false. That the q-values are generally smaller than their corresponding p-values is an indication that $S > m_0$ since the p- and q-values are measures of $E(F/m_0)$ and $E(F/S)$ respectively.

Consequently, when we use 10% or 5% rejection thresholds for the p-values the corresponding q-values indicate that we expect at most $E(F_{10\%}/S_{10\%})$ = 4.9% or $E(F_{5\%}/S_{5\%})$ = 3.9% of the rejections correspond to false discoveries. Since the implied expected number of

false positives (i.e. $E(F_{10\%}) = 4.9\% \cdot 154 \cong 8$ and $E(F_{5\%}) = 3.9\% \cdot 94 \cong 4$; see the lower panel of Figure 2 for more detail) under the family-wise null is much smaller than the number of rejections indicated by 10% and 5% p-value rejection thresholds, and since the expected number of false discoveries is fairly small as a proportion of those rejections (4.9% – 3.9%) we use both 10% and 5% p-value rejection thresholds as rules for detecting predictive ability in the structural models – despite the existence of multiple testing.

Tests of Equal Forecast Accuracy

Table 2 presents the results for the significance of the MSE-t statistics. These statistics provide evidence of short-, medium- and long-horizon predictability. For example, model 1 provides evidence of 1- and 2-quarter predictability for the Swiss franc with both p- and q-values less than 5%. At the 4-quarter horizon the corresponding p-value is less than 10% and has a q-value less than 5%. For the German mark there is evidence of predictability at the 1- and 2-quarter horizons and for the Japanese yen at most horizons. This pattern holds consistently across all models. The evidence of predictability using this test statistic is very weak for the Canadian dollar. This seems somewhat surprising since among all of the RMSE ratios in Table 1, the largest values are obtained at the 12-quarter horizon for Canada and many of the values at this and other forecast horizons are statistically significant.

Table 3 presents the results for the significance of the MSE-F test statistic. If we compare Tables 2 and 3, in all of the cases in which the MSE-t has a p-value less than 10%, the MSE-F statistic does so as well. The same holds for p-values of less than 5%. In some cases however, the MSE-F has a p-value less than 5% or 10% when the MSE-t does not. Recall that we motivated the use of the MSE-F statistic by arguing that it has greater power than the standard

MSE-t statistic. Our empirical findings are consistent with this. There are 10 (11) instances in which the MSE-F has a p-value less than 10% (5%) but the MSE-t does not. In most other cases, we also find the p-values are much lower for the MSE-F statistics than for the MSE-t statistics. Of particular interest are the 12-quarter forecasts for Canada where the p-values for the MSE-t are around 77% but they are only slightly over our 10% cut-off for the MSE-F statistics.

In light of the superior power of the MSE-F statistic, it is not surprising that the MSE-F indicates increased predictive ability for each model and most currencies. In particular there is stronger evidence of predictive ability at the 4-quarter horizon. For each of the German mark, Swiss franc and Canadian dollar we now find predictive ability at the 4-quarter horizon across all models. There are no significant changes in predictive ability for the Japanese yen.

Tests of Forecast Encompassing

Moving to the tests of encompassing, Table 4 presents the evidence of predictability associated with the ENC-t statistic. The evidence of predictive ability here is closely related to what we found using the MSE-t statistic in Table 2. The main difference is that we find slightly fewer significant results. The largest difference between Tables 2 and 4 occurs for the Japanese yen. The MSE-t statistic indicates predictive ability across a wide variety of horizons while the ENC-t statistic would only indicate predictive ability for the 1- and 2- quarter horizons. Again, the MSE-F statistic signals significantly more predictive ability than does the ENC-t statistic and as was the case for the MSE-t statistic, this is particularly true for the Canadian dollar.

Table 5 presents the evidence of predictive ability associated with the ENC-F test statistic. If we compare Tables 4 and 5, in all but two instances when the ENC-t rejects the null at the 10% level, the ENC-F continues to do so. In both of the cases for which this did not occur, the ENC-F would have been significant at the 11% level or better. In some cases however, the ENC-F

rejects the null when the ENC-t does not. There are 20 (8) instances in which the ENC-F has a p-value less than 10% (5%) but the ENC-t does not.

As was the case for the MSE-F test, we observe a sharp increase in the predictive ability signaled by the ENC-F statistic relative to its' t-type companion. For example whereas the ENC-t statistic did not indicate any predictability of the Canadian dollar, we now observe evidence of predictability at a wide variety of horizons across all models. It is also interesting to reconsider the Canadian dollar at the 12-quarter forecast horizon but this time noting the difference in the p-values for the ENC-t and ENC-F statistics rather than the MSE statistics. For the ENC-t statistic the p-values are over 90% but they are just slightly over 10% for the ENC-F statistic. Similarly, recall that the ENC-t statistic indicated only 1- and 2-quarter predictive ability of the German mark, Japanese yen and Swiss franc. The ENC-F statistic now provides evidence of 1-, 2- and 4-quarter predictability across most models for each of these currencies.

The final test of forecasting encompassing we use is that of Chao, Corradi and Swanson (2001). In Table 6 we see that this test indicates less predictive ability for our models than did our earlier tests. At the 5% level we find evidence of predictive ability at the 12-quarter horizon for the Canadian dollar and Swiss franc. Since few of the other tests indicate predictive ability for the Canadian dollar despite the large RMSE ratios, it may be that this test has significantly better power in particular directions relative to the other tests (see Corradi and Swanson (2002) for a related discussion). We find a large number of CCS-statistics with p-values between 10% and 15%. Even though not statistically significant, this suggests some predictive ability for the forecasts at the 8-quarter horizon for both the Canadian dollar and Swiss Franc and at the 1- and 2-quarter horizons for the Japanese yen. Consequently this test provides a valuable new perspective on the predictive ability of our models.

5. Conclusion

Our analysis suggests that detecting predictability in exchange rates using long-horizon regressions can be strongly influenced by the choice of test statistic and the manner in which it is employed. In particular we find clear empirical evidence that the standard t-type test of equal forecast accuracy that is used in Meese and Rogoff (1988), Mark (1995) and Kilian (1999) and Cheung, Chinn and Pascual (2002) indicates significantly less predictability than its' F-type equivalent. A comparable relationship is established between t- and F-type tests of forecast encompassing. Although analytical and Monte Carlo evidence of such relationships are established in Clark and McCracken (2001, 2003a, b) it is clearly more important that such a relationship is established empirically. It is therefore comforting to find that these conclusions seem to hold irrespective of the particular currency or model being considered.

Our results yield several implications for researchers dealing with financial time series. For those in International Finance it adds further evidence that structural exchange rate models do exhibit an ability to predict exchange rates. Similar to other studies our evidence is consistent with there being more short-term predictability in exchange rates and our results are relatively insensitive to the choice of model. More generally it suggests that these new, more powerful, test statistics may be useful for ascertaining whether a particular financial variable with putative predictive content indeed can be used to improve forecast accuracy.

References

Benjamini, Y. and Y. Hochberg. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society, Series B, 57, 289-300.

Benjamini, Y. and D. Yekutieli. (2001). "The Control of the False Discovery Rate in Multiple Testing Under Dependency." The Annals of Statistics, 29, 1165-1188.

Berben, R.P. and D. van Dijk. (1998). "Does the Absence of Cointegration Explain the Typical Findings in Long Horizon Regressions?" Econometric Institute Report 145, Erasmus University.

Berkowitz, J. and L. Giorgianni. (2001). "Long-Horizon Exchange Rate Predictability?" Review of Economics and Statistics, 83, 81-91.

Bilson, J. (1978). "The Current Experience with Floating Exchange Rates: An Appraisal of the Monetary Approach." American Economic Review, 68, 392-97.

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

Cheung, Y., M. Chinn and A. Pascual. (2002). "Empirical Exchange Rate Models of the Nineties: Are Any Fit to Survive?." Journal of International Money and Finance, forthcoming.

Chinn, M. and R. Meese. (1995). "Banking On Currency Forecasts: How Predictable is Change in Money?" Journal of International Economics, 38, 161-178.

Chong, Y.Y. and D.F. Hendry. (1986). "Econometric Evaluation of Linear Macroeconomic Models." Review of Economic Studies, 53, 671-90.

Clarida, R., L. Sarno, M. Taylor and G. Valente. (2003). "The Out-of-Sample Success of Term Structure Models as Exchange Rate Predictors: A Step Beyond." Journal of International Economics, 60, 61-83.

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

Clark, T.E. and M.W. McCracken. (2003a). "Evaluating Long Horizon Forecasts," manuscript, Federal Reserve Bank of Kansas City and University of Missouri-Columbia.

Clark, T.E. and M.W. McCracken. (2003b). "The Power of Tests of Predictive Ability in the Presence of Structural Change', Journal of Econometrics, forthcoming.

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

Diebold, F.X. and R.S. Mariano. (1995). "Comparing Predictive Accuracy," Journal of Business and Economic Statistics, 13, 253-63.

Faust, J., J. Rogers and J. Wright. (2003). "Exchange Rate Forecasting: The Errors We've Made." Journal of International Economics, 60, 35-59.

Frankel, J. and A. Rose. (1995). "Empirical Research on Nominal Exchange Rates." in Handbook of International Economics Vol. III, edited by G. Grossman and K. Rogoff, pp. 1689-1729. Elsevier Science.

Frenkel, J. (1976). "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence." Scandinavian Journal of Economics, 78, 200-224.

Harvey, D.I., S.J. Leybourne and P. Newbold. (1998). "Tests for Forecast Encompassing." Journal of Business and Economic Statistics, 16, 254-59.

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

Kilian, L. and M. Taylor. (2003). "Why is it so difficult to beat the random walk forecast of exchange rates?" Journal of International Economics, 60, 85-107.

Levich, R. (1985). "Empirical Studies of Exchange Rates: Price Behavior, Rate Determination and Market Efficiency." in Handbook of International Economics Vol. II, edited by R.W. Jones and P.B. Kenen, pp. 980-1040. North Holland.

MacDonald, R. and I. Marsh. (1997). "On Fundamentals and Exchange Rates: A Casselian Perspective." Review of Economics and Statistics, 79, 655-664.

Mark, N. (1995). "Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability." American Economic Review, 85, 201-218.

Mark, N. and D. Sul. (2002). "Asymptotic Power Advantages of Long-Horizon Regressions." manuscript, Ohio State University.

McCracken, M.W. (2000). "Asymptotics for Out-of-Sample Tests of Causality." manuscript, University of Missouri-Columbia.

Meese, R.A. and K. Rogoff. (1983a). "Empirical Exchange Rate Models of the Seventies: Do They Fit Out-of-Sample?" Journal of International Economics, 14, 3-24.

Meese, R and K. Rogoff. (1983b). "The Out-of-Sample Failure of Empirical Exchange Rate Models." in Exchange Rates and International Macroeconomics, edited by J. Frenkel, pp. 67-105. Chicago: University of Chicago Press.

Meese, R.A. and K. Rogoff. (1988). "Was it Real? The Exchange Rate-Interest Differential Relation over the Modern Floating-Rate Period." Journal of Finance, 43, 933-948.

Mussa, M. (1976). "The Exchange Rate, the Balance of Payments and Monetary and Fiscal Policy under a Regime of Controlled Floating." Scandinavian Journal of Economics, 78, 229-48.

Neely, C. and L. Sarno. (2002). "How Well Do Monetary Fundamentals Forecast Exchange Rates?" Review Federal Reserve Bank of Saint Louis, 84, 51-74.

Newey, W. K. and K.D. West. (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." Econometrica, 55, 703-08.

Rapach, D. and M. Wohar. (2002). "Testing the Monetary Model of Exchange Rate Determination: New Evidence from a Century of Data." Journal of International Economics, 58, 359-385.

Rossi, B. (2002). "Testing Long-Horizon Predictive Ability with High Persistence and the Meese-Rogoff Puzzle." manuscript, Duke University.

Storey, J.D. (2002). "A Direct Approach to False Discovery Rates." Journal of the Royal Statistical Society, Series B, 64, 479-498.

Storey, J.D. (2003). "The Positive False Discovery Rate: A Bayesian Interpretation and the q-value." The Annals of Statistics, 31, 2013-2035.

Storey, J.D., J.E. Taylor and D. Siegmund. (2003). "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach." Journal of the Royal Statistical Society, Series B, 66, 187-205.

Storey J.D. and R. Tibshirani. (2001). “Estimating false discovery rates under dependence with applications to DNA microarrays.” Technical Report 2001-28, Department of Statistics, Stanford University.

Storey, J.D. and R. Tibshirani. (2003). “Statistical Significance for Genomewide Studies.” Proceedings of the National Academy of Sciences, 100, 9440-9445.

Taylor, M., D. Peel and L. Sarno. (2001). “Non-linear Mean-Reversion in Real Exchange Rates: Towards a Solution to the Purchasing Power Parity Puzzles.” International Economic Review, 42, 1015-1042.

West, K.D. (1996). “Asymptotic Inference About Predictive Ability.” Econometrica, 64, 1067-1084.

Table 1: Sample Forecasting Results

		Random Walk	Model 1	Model 2	Model 3	Model 4
		RMSE	Ratio			
Germany	1	24.50	1.02	1.02	1.02	1.02
	2	34.38	1.06	1.06	1.06	1.06
	4	43.66	1.12	1.13	1.12	1.12
	8	53.56	0.89	0.91	0.89	0.89
	12	41.37	0.53	0.55	0.53	0.53
Canada	1	7.95	1.01	1.01	1.01	1.01
	2	10.77	1.02	1.03	1.03	1.03
	4	15.85	1.13	1.14	1.14	1.14
	8	26.19	1.76	1.78	1.76	1.75
	12	34.61	3.57	3.60	3.50	3.41
Japan	1	25.66	1.04	1.04	1.04	1.04
	2	35.93	1.07	1.08	1.08	1.07
	4	45.85	1.09	1.10	1.09	1.09
	8	76.02	1.09	1.12	1.11	1.11
	12	83.33	0.93	0.95	0.94	0.94
Switzerland	1	27.32	1.05	1.05	1.05	1.05
	2	38.71	1.10	1.10	1.10	1.10
	4	49.40	1.23	1.22	1.24	1.23
	8	63.30	1.39	1.34	1.38	1.37
	12	51.85	1.07	1.01	1.03	1.00

Notes: (a) Values are the Root Mean Squared Error (RMSE) for the random walk with drift and its' ratio with the RMSE for each of the structural models. A value greater than one favors the structural model.
(b) Forecasts from the structural models use fundamentals based on parameter from previous studies such as Mark (1995), Kilian (1999) and Cheung, Chinn and Pascual (2002).
(c) Initial model estimates use quarterly data over the period 1973 to 1989 while the out-of-sample forecasts are over the period 1990 to 1998.
(d) Where models 1 to 4 are the monetary model, flexible price monetary model (Frenkel-Bilson), sticky price monetary model (Dornbusch-Frankel) and sticky-price asset model (Hooper-Morton), respectively.

Table 2: p- and q-values for the MSE-t statistic

	τ	Model 1		Model 2		Model 3		Model 4	
		p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Canada	1	0.1865	0.0588	0.1865	0.0588	0.1940	0.0606	0.1960	0.0609
	2	0.2625	0.0744	0.2695	0.0749	0.2610	0.0744	0.2750	0.0751
	4	0.2255	0.0679	0.2190	0.0662	0.2140	0.0653	0.2080	0.0642
	8	0.3505	0.0840	0.3480	0.0838	0.3515	0.0840	0.3630	0.0862
	12	0.7770	0.1539	0.7635	0.1524	0.7760	0.1539	0.7740	0.1539
Germany	1	0.0455	0.0388	0.0390	0.0388	0.0400	0.0388	0.0450	0.0388
	2	0.0690	0.0426	0.0580	0.0388	0.0570	0.0388	0.0600	0.0393
	4	0.1495	0.0514	0.1405	0.0501	0.1515	0.0514	0.1520	0.0514
	8	0.2670	0.0747	0.2385	0.0707	0.2755	0.0751	0.2790	0.0758
	12	0.4850	0.1034	0.4930	0.1045	0.4965	0.1048	0.5050	0.1053
Japan	1	0.0085	0.0388	0.0085	0.0388	0.0065	0.0388	0.0070	0.0388
	2	0.0520	0.0388	0.0450	0.0388	0.0415	0.0388	0.0330	0.0388
	4	0.2095	0.0644	0.1875	0.0588	0.1855	0.0588	0.1965	0.0609
	8	0.0355	0.0388	0.0330	0.0388	0.0505	0.0388	0.0410	0.0388
	12	0.0355	0.0388	0.0320	0.0388	0.0395	0.0388	0.0345	0.0388
Switzerland	1	0.0035	0.0335	0.0095	0.0388	0.0035	0.0335	0.0065	0.0388
	2	0.0145	0.0388	0.0190	0.0388	0.0135	0.0388	0.0210	0.0388
	4	0.0780	0.0444	0.1090	0.0483	0.0735	0.0431	0.0860	0.0453
	8	0.1585	0.0526	0.1570	0.0525	0.1575	0.0525	0.1475	0.0509
	12	0.4720	0.1022	0.5010	0.1049	0.4745	0.1022	0.4655	0.1019

Notes: (a) Values are the p- and q-values associated with tests of predictive ability using the MSE-t statistic. The p-values are calculated using a bootstrap based on Kilian (1999) while the q-values are calculated using an algorithm due to Storey (2003). Whereas the p-value provides a measure of the rate at which null hypotheses are rejected, the q-value provides a measure of the rate at which rejected hypotheses satisfy the null hypothesis.
(b) See notes (b) to (d) from Table 1.

Table 3: p- and q-values for the MSE-F statistic

	τ	Model 1		Model 2		Model 3		Model 4	
		p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Canada	1	0.1095	0.0483	0.0950	0.0477	0.1085	0.0483	0.1050	0.0483
	2	0.1290	0.0487	0.1275	0.0487	0.1225	0.0487	0.1290	0.0487
	4	0.0560	0.0388	0.0350	0.0388	0.0505	0.0388	0.0405	0.0388
	8	0.1350	0.0487	0.1295	0.0487	0.1470	0.0509	0.1435	0.0509
	12	0.1025	0.0483	0.1065	0.0483	0.1215	0.0487	0.1325	0.0487
Germany	1	0.0500	0.0388	0.0465	0.0388	0.0475	0.0388	0.0490	0.0388
	2	0.0630	0.0405	0.0485	0.0388	0.0560	0.0388	0.0565	0.0388
	4	0.0895	0.0461	0.0830	0.0451	0.0840	0.0451	0.0980	0.0483
	8	0.2990	0.0776	0.2730	0.0750	0.2985	0.0776	0.3085	0.0797
	12	0.5200	0.1069	0.5105	0.1059	0.5160	0.1064	0.5365	0.1094
Japan	1	0.0140	0.0388	0.0130	0.0388	0.0125	0.0388	0.0185	0.0388
	2	0.0415	0.0388	0.0375	0.0388	0.0355	0.0388	0.0340	0.0388
	4	0.1200	0.0487	0.1080	0.0483	0.1060	0.0483	0.1205	0.0487
	8	0.0310	0.0388	0.0285	0.0388	0.0405	0.0388	0.0370	0.0388
	12	0.0395	0.0388	0.0340	0.0388	0.0425	0.0388	0.0390	0.0388
Switzerland	1	0.0210	0.0388	0.0220	0.0388	0.0215	0.0388	0.0175	0.0388
	2	0.0220	0.0388	0.0225	0.0388	0.0275	0.0388	0.0220	0.0388
	4	0.0255	0.0388	0.0280	0.0388	0.0225	0.0388	0.0230	0.0388
	8	0.1310	0.0487	0.1325	0.0487	0.1295	0.0487	0.1230	0.0487
	12	0.4740	0.1022	0.5010	0.1049	0.4735	0.1022	0.4640	0.1018

Notes: (a) Values are the p- and q-values associated with tests of predictive ability using the MSE-F statistic. The p-values are calculated using a bootstrap based on Kilian (1999) while the q-values are calculated using an algorithm due to Storey (2003).
(b) See notes (b) to (d) from Table 1.

Table 4: p- and q-values for the ENC-t statistic

	τ	Model 1		Model 2		Model 3		Model 4	
		p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Canada	1	0.1130	0.0483	0.1070	0.0483	0.1195	0.0487	0.1200	0.0487
	2	0.1675	0.0545	0.1825	0.0582	0.1680	0.0545	0.1795	0.0575
	4	0.1595	0.0527	0.1520	0.0514	0.1525	0.0514	0.1530	0.0514
	8	0.4745	0.1022	0.4720	0.1022	0.4490	0.1000	0.4450	0.0994
	12	0.9200	0.1749	0.9160	0.1749	0.9125	0.1748	0.9190	0.1749
Germany	1	0.0735	0.0431	0.0690	0.0426	0.0705	0.0426	0.0620	0.0402
	2	0.0830	0.0451	0.0705	0.0426	0.0705	0.0426	0.0700	0.0426
	4	0.1330	0.0487	0.1310	0.0487	0.1355	0.0487	0.1360	0.0487
	8	0.3805	0.0895	0.3895	0.0908	0.3835	0.0897	0.3715	0.0877
	12	0.4210	0.0966	0.4010	0.0930	0.4115	0.0947	0.4015	0.0930
Japan	1	0.0410	0.0388	0.0265	0.0388	0.0425	0.0388	0.0365	0.0388
	2	0.0990	0.0483	0.0890	0.0461	0.0935	0.0472	0.0850	0.0451
	4	0.3200	0.0806	0.2940	0.0774	0.3250	0.0813	0.3170	0.0801
	8	0.8870	0.1712	0.8820	0.1712	0.8770	0.1710	0.8840	0.1712
	12	0.9775	0.1840	0.9735	0.1837	0.9720	0.1837	0.9660	0.1832
Switzerland	1	0.0015	0.0205	0.0045	0.0339	0.0015	0.0205	0.0040	0.0335
	2	0.0215	0.0388	0.0180	0.0388	0.0155	0.0388	0.0190	0.0388
	4	0.1135	0.0483	0.1340	0.0487	0.1090	0.0483	0.1070	0.0483
	8	0.3385	0.0829	0.3405	0.0829	0.3270	0.0815	0.2980	0.0776
	12	0.3470	0.0838	0.3570	0.0850	0.3675	0.0870	0.4065	0.0939

Notes: (a) Values are the p- and q-values associated with tests of predictive ability using the ENC-t statistic. The p-values are calculated using a bootstrap based on Kilian (1999) while the q-values are calculated using an algorithm due to Storey (2003).
(b) See notes (b) to (d) from Table 1.

Table 5: p- and q-values for the ENC-F statistic

	τ	Model 1		Model 2		Model 3		Model 4	
		p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Canada	1	0.0570	0.0388	0.0540	0.0388	0.0580	0.0388	0.0505	0.0388
	2	0.0595	0.0393	0.0530	0.0388	0.0465	0.0388	0.0470	0.0388
	4	0.0540	0.0388	0.0315	0.0388	0.0470	0.0388	0.0340	0.0388
	8	0.1675	0.0545	0.1775	0.0572	0.1775	0.0572	0.1675	0.0545
	12	0.1065	0.0483	0.1150	0.0486	0.1215	0.0487	0.1290	0.0487
Germany	1	0.1040	0.0483	0.0965	0.0481	0.1000	0.0483	0.0930	0.0472
	2	0.1005	0.0483	0.0780	0.0444	0.0845	0.0451	0.0835	0.0451
	4	0.0925	0.0472	0.0725	0.0430	0.0805	0.0451	0.0870	0.0455
	8	0.3240	0.0813	0.2870	0.0774	0.3160	0.0801	0.3150	0.0801
	12	0.5140	0.1063	0.4785	0.1026	0.5065	0.1053	0.4965	0.1048
Japan	1	0.0765	0.0443	0.0545	0.0388	0.0700	0.0426	0.0705	0.0426
	2	0.0835	0.0451	0.0695	0.0426	0.0810	0.0451	0.0785	0.0444
	4	0.1105	0.0483	0.1010	0.0483	0.1045	0.0483	0.1160	0.0487
	8	0.5510	0.1110	0.5215	0.1070	0.5275	0.1079	0.5450	0.1106
	12	0.9010	0.1730	0.8730	0.1707	0.8910	0.1715	0.8850	0.1712
Switzerland	1	0.0465	0.0388	0.0430	0.0388	0.0450	0.0388	0.0430	0.0388
	2	0.0480	0.0388	0.0540	0.0388	0.0575	0.0388	0.0560	0.0388
	4	0.0510	0.0388	0.0475	0.0388	0.0465	0.0388	0.0420	0.0388
	8	0.2610	0.0744	0.2595	0.0744	0.2510	0.0741	0.2350	0.0699
	12	0.5530	0.1110	0.5545	0.1110	0.5530	0.1110	0.5385	0.1096

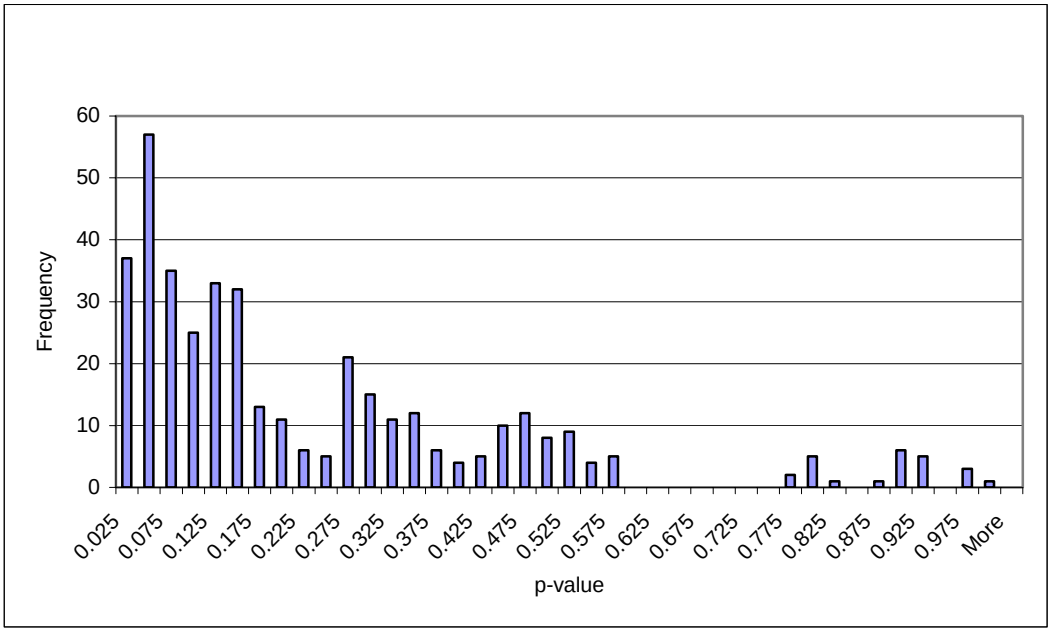
Notes: (a) Values are the p- and q-values associated with tests of predictive ability using the ENC-F statistic. The p-values are calculated using a bootstrap based on Kilian (1999) while the q-values are calculated using an algorithm due to Storey (2003).
(b) see notes (b) to (d) from Table 1.

Table 6: p- and q-values for the CCS statistic

	τ	Model 1		Model 2		Model 3		Model 4	
		p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Canada	1	0.2945	0.0774	0.2943	0.0774	0.2943	0.0774	0.2940	0.0774
	2	0.3125	0.0797	0.3123	0.0797	0.3123	0.0797	0.3120	0.0797
	4	0.4295	0.0974	0.4292	0.0974	0.4290	0.0974	0.4289	0.0974
	8	0.1136	0.0483	0.1135	0.0483	0.1134	0.0483	0.1133	0.0483
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Germany	1	0.3281	0.0815	0.3364	0.0829	0.3483	0.0838	0.3819	0.0895
	2	0.2142	0.0653	0.2185	0.0662	0.2272	0.0682	0.2611	0.0744
	4	0.4835	0.1034	0.4873	0.1036	0.4989	0.1049	0.5521	0.1110
	8	0.7838	0.1545	0.7823	0.1545	0.7870	0.1547	0.8237	0.1615
	12	0.2520	0.0741	0.2705	0.0749	0.2538	0.0744	0.2339	0.0699
Japan	1	0.1345	0.0487	0.1323	0.0487	0.1314	0.0487	0.1327	0.0487
	2	0.1473	0.0509	0.1457	0.0509	0.1450	0.0509	0.1465	0.0509
	4	0.2949	0.0774	0.2906	0.0774	0.2916	0.0774	0.2940	0.0774
	8	0.2727	0.0750	0.2645	0.0744	0.2649	0.0744	0.2682	0.0748
	12	0.4617	0.1016	0.4593	0.1014	0.4579	0.1014	0.4580	0.1014
Switzerland	1	0.4393	0.0984	0.4374	0.0984	0.4380	0.0984	0.4385	0.0984
	2	0.3415	0.0829	0.3405	0.0829	0.3408	0.0829	0.3410	0.0829
	4	0.2632	0.0744	0.2626	0.0744	0.2626	0.0744	0.2622	0.0744
	8	0.1316	0.0487	0.1302	0.0487	0.1305	0.0487	0.1301	0.0487
	12	0.0289	0.0388	0.0281	0.0388	0.0287	0.0388	0.0289	0.0388

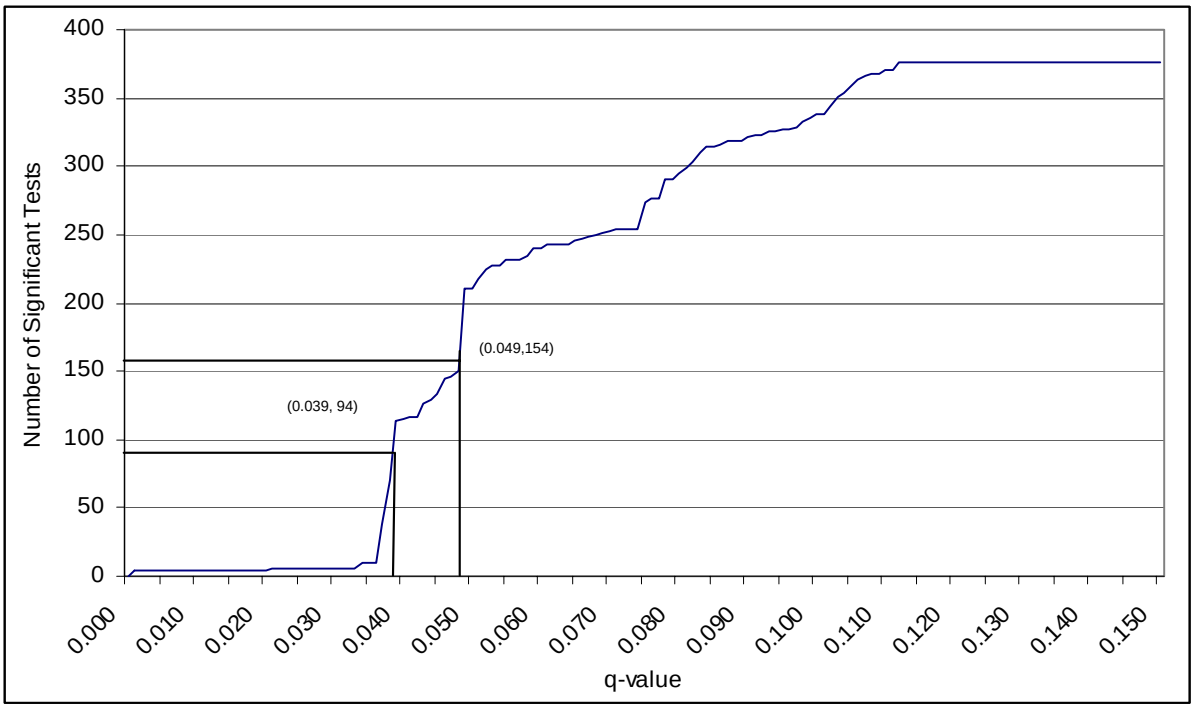
Notes: (a) Values are the p- and q-values associated with tests of predictive ability using the CCS statistic. The p-values are calculated using the upper tail of the chi-square(1) distribution while the q-values are calculated using an algorithm due to Storey (2003).
(b) see notes (b) to (d) from Table 1.

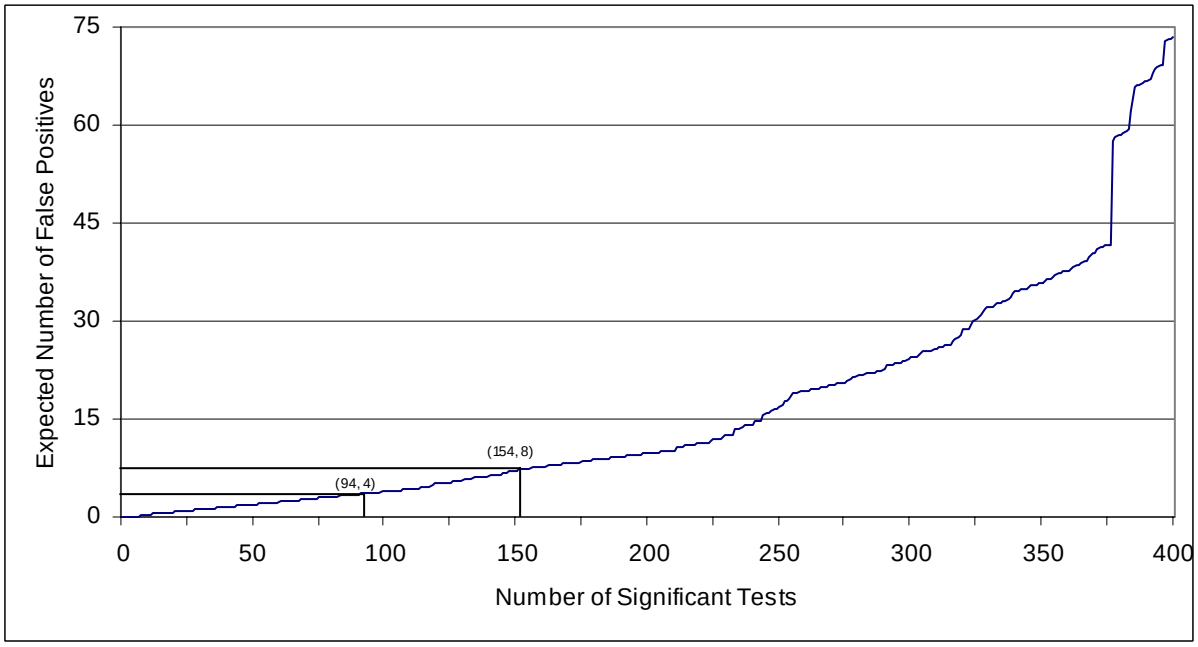
Figure 1: Histogram of All p-values



Notes: (a) This chart provides a histogram of all 400 p-values generated from the permutations of 5 test statistics, 5 forecast horizons, 4 bilateral exchange rates and 4 models.
(b) Bins are in increments of .025.
(c) The number of p-values less than .05 and .10 are 94 and 154 respectively.

Figure 2: Plots of q-values by S and E(F)





Notes: (a) These charts describe the q-values corresponding to all 400 p-values generated from the permutations of 5 test statistics, 5 forecast horizons, 4 bilateral exchange rates and 4 models.
(b) The figures are: a) the number of significant tests (tests in which we reject the null of no predictive ability) versus the respective q-value, and b) the expected number of false positives versus the number of significant tests.

* We would like to thank Todd Clark, Lutz Kilian, Mark Wohar, seminar participants at the 2001 MEG meetings and the Bank of Canada as well as three anonymous referees and the editor for helpful comments. All remaining errors are our own.

[1] Recent studies such as Cheung, Chinn and Pascual (2002) and Rossi (2002) concentrate on different characteristics of the problem and also raise concerns regarding the use of standard t-type tests of predictive ability.

[2] For a detailed discussion of some of these and other related issues see Neely and Sarno (2002).

[3] Throughout, the notation '*' signifies that the variable relates to the foreign country.

[4] This description is drawn largely from Storey and Tibshirani (2003).

[5] Details on these models can be found in Levich (1985), and Frankel and Rose (1995).

[6] These assumed values for the coefficients in the model are taken from studies such as Mark (1995), Kilian (1999) and Cheung, Chinn and Pascual (2002).

[7] We also performed this analysis with estimated coefficients, but the results were qualitatively similar. Using preset values also facilitates comparison with much of the existing literature .

[8] This is estimated separately for each τ, but we suppress the dependence on τ for simplicity.

[9] We consider several other sample splits. As in Faust, Rogers and Wright (2003), Cheung, Chinn and Pascual (2002) and many other studies, we found that the forecasting ability of the different models were impacted by the choice of sample period. However, the relative performance of the different statistics remained consistent across our alternative sample periods. Consequently, we only present the results for the split similar to Kilian (1999).

[10] We also considered using 20 lags as in Kilian (1999) but obtained similar results.

[11] In unreported work, we construct tests of zero mean prediction error over the out-of-sample period. For all models, horizons and exchange rates we fail to reject this null at the 10% level.