NBER WORKING PAPER SERIES

THE CONTINUING PUZZLE OF SHORT HORIZON EXCHANGE RATE FORECASTING

Kenneth S. Rogoff
Vania Stavrakeva

Working Paper 14071
http://www.nber.org/papers/w14071

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2008

We are extremely grateful to Barbara Rossi, Guido Imbens, Brent Neiman, William Dickens, Konstantin Styrin, Charles Engel, David Papell, Todd Clark, Kenneth West, Jinzhu Chen and the participants at the Harvard International Lunch Seminar for their excellent suggestions and comments. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

The Continuing Puzzle of Short Horizon Exchange Rate Forecasting
Kenneth S. Rogoff and Vania Stavrakeva
NBER Working Paper No. 14071
June 2008, Revised August 2008
JEL No. C52,C53,F31,F47

ABSTRACT

Are structural models getting closer to being able to forecast exchange rates at short horizons? Here we argue that misinterpretation of some new out-of-sample tests for nested models, over-reliance on asymptotic test statistics, and failure to sufficiently check robustness to alternative time windows have led many studies to overstate even the relatively thin positive results that have been found. We find that by allowing for common cross-country shocks in our panel forecasting specification, we are able to generate some improvement, but even that improvement is not entirely robust to the forecast window, and much of the gain appears to come from non-structural rather than structural factors.

Kenneth S. Rogoff
Thomas D Cabot Professor of Public Policy
Economics Department
Harvard University
Littauer Center 232
Cambridge, MA 02138-3001
and NBER
krogoff@harvard.edu

Vania Stavrakeva
The Brookings Institution
1775 Massachusetts Avenue, N.W.
Washington, DC 20036
vania.stavrakeva@gmail.com

Introduction

Understanding the connection between exchange rates and macroeconomic fundamentals has been one of the central challenges in international macroeconomics since the start of the modern floating exchange rate era in the early 1970s. Although exchange rates are indeed an asset price, and, therefore, highly volatile, they also reflect basic macroeconomic fundamentals such as interest rates, purchasing power, and trade balances. As such, international economists have long held out hope they could explain exchange rates better than, say, finance economists can explain the absolute level of stock prices. If so, the results would be of enormous help to policy-makers including, for example, central bankers who might worry about the effect of monetary policy on exchange rates.

Unfortunately, in practice, the performance of structural exchange rate models has been frustratingly disappointing. As first shown by Meese and Rogoff (1983a), models that perform well in-sample seldom do so out-of-sample. Although one can find some forecasting power at horizons of two to four years (e.g., Meese and Rogoff, 1983b, Mark, 1995 or Engel, Mark and West, 2007), attempts to forecast at more policy-relevant horizons of one month to one year have been far less successful.[1]

Indeed, until recently, there had been surprisingly little progress despite hundreds of studies using a plethora of techniques (see Cheung, Chinn and Pascual, 2003, for a survey). Lately, however, the literature has experienced a new life. A growing number of papers have been reporting somewhat more positive short-term forecasting results by implementing panel forecast methods, innovative estimation procedures, more powerful out-of-sample test statistics and new structural models. These include influential papers by Gourinchas and Rey (2007), Engel, Mark and West (2007) and Molodtsova and Papell (2008) along with many other notable studies.[2] This paper re-examines the new evidence and considers a number of variations and refinements.

We conclude that despite notable methodological improvements, the euphoria has been exaggerated by *misinterpretation of some newer out-of-sample test statistics for nested models*, *over-reliance on asymptotic out-of-sample test statistics* and *failure to check for robustness to the time period sampled*.

Our examination of the most popular exchange rate forecasting structural models and specifications leads us to conclude that one of the sources of the overly optimistic results is the failure to check robustness with respect to alternative out-of-sample test statistics. In the presence of

[1] Further research is required to determine the robustness of the long-horizon forecastability results with respect to using different sub-samples. For instance, Mark's (1995) results do not hold when one updates his sample (see Kilian, 1999).

[2] See also Rapach and Wohar (2002), Rossi (2006), Groen (2005, 2007), Cerra and Saxena (2008), Ardic et al. (2008), Molodtsova, Nikolsko-Rzhevskyy and Papell (2007, 2008) and Sellin (2006).

forecast bias[3], the new tests for nested models[4] cannot be always interpreted as minimum mean square forecast error tests.[5] As we show, in certain cases, this is a first-order problem. Furthermore, while new asymptotic out-of-sample tests such as the Clark-West are attractive due to their simplicity, bootstrapped out-of-sample tests remain more powerful and better sized.[6] Finally, even if the results remain statistically significant if one considers alternative out-of-sample test statistics, all of the structural models and specifications we review fail to produce robust forecasts over different sample periods, implying that in one period the random walk is a better forecaster and in another the structural model outperforms the random walk. In such cases, even if a structural model performs well during the most recent period of time, there is no guarantee that the relationship will be preserved in the future.

The paper is organized in the following way. Section 1 sets out our criteria for what constitutes a "good" forecast — a forecast with a mean-square forecast error smaller than the mean-square forecast error of the driftless random walk, and with robust out-of-sample test statistics over different forecast windows.[7] In section 2, we introduce the out-of-sample tests we consider — the asymptotic Clark-West and the bootstrapped Diebold-Mariano/West, Theil's U, Clark-West and Clark-McCracken test statistics. We discuss the differences between the alternative test statistics, the most important of which is that in cases of forecast bias, the new nested model tests should be interpreted as testing against the null hypothesis that the true model is a random walk, rather than as asking whether a random walk has a lower mean-square forecast error than the structural model (which is what the older Theil's U and Diebold-Mariano/West statistics test). These turn out to be quite different questions, although we also show that the newer nested model statistics can point to cases where it may be possible to improve on the random walk forecast by using it in combination with the structural model forecast. Nevertheless, finding an endogenous optimal combination may be a significant obstacle.

Section 3 tests the robustness of the apparent best results of the literature on short-horizon forecasting with respect to using alternative out-of-sample test statistics. The main studies reviewed are Gourinchas and Rey (2007) — an external balance model; Molodtsova and Papell (2008) — a heterogeneous symmetric Taylor rule model with smoothing; and Engel, Mark and

[3]In this paper we are concerned only with "scale" bias as opposed to "location" bias. In other words, our result refers only to the cases where the forecast systematically over or under-predicts the observed value by a certain percent (see Holden and Peel (1989) for a distinction between the two types of bias). For a general definition of forecast bias see Marcellino (2000), pp. 534.

[4]These tests include the Clark and West (2006, 2007) and the Clark and McCracken (2001, 2005).

[5]This is a problem with both the asymptotic and the bootstrapped Clark-West and Clark-McCracken

[6]The advance of the literature on time series bootstrapping and the increase of computational power have made the bootstrap an increasingly attractive alternative to asymptotic inference (see Berkowitz and Kilian,1996, Kilian, 1999, Mark and Sul, 2001, MacKinnon, 2002, Brownstone and Valletta, 2001, and Politis and White, 2004). For a detailed discussion of how the bootstrap can provide a significant improvement over asymptotic inference see Li and Maddala (1997).

[7]"Forecast window" refers to the part of the sample for which forecasts are calculated. For example, if we have a sample of 120 quarters and the first forecast is based on 30 quarters, then the forecast window is 90 quarters.

West (2007) — the monetary model. We conclude that in certain cases the popular Clark-West and Clark-McCracken test statistics are highly significant while the bootstrapped Theil's U and Diebold-Mariano/West are not which we attribute to the presence of forecast bias. Furthermore, in a couple of cases, the asymptotic Clark-West incorrectly chooses the structural model forecast over the random walk forecast for a different reason – the asymptotic Clark-West test seems to be oversized. In section 4, we explore the robustness of the results of these same studies with respect to different forecast windows using a graphic approach which illustrates how the significance of the results is affected by perturbing the sample. We find that even those results that are robust to alternative out-of-sample test statistics are not robust when the forecaster considers alternative samples, with the external balance model of Gourinchas and Rey (2007) performing somewhat better than the rest of the specifications considered. This point strongly reinforces our conclusion from section 3 that the results of the new models and specifications are not very robust.

Therefore, we attempt to improve upon existing panel specifications in section 5 by taking into account persistent cross-country shocks using purchasing power parity as a fundamental. Similarly to the results in section 3, our results point to a discrepancy between the old out-of-sample test statistics and the new out-of-sample tests for nested models. In section 6 we present empirical evidence of how one can improve upon our results from section 5 by correctly interpreting the new nested model tests and combining the structural model forecast and the random walk forecast. At first look, our results are not worse than the most prominent results of other existing short-horizon forecasting studies. Nevertheless, the fact that so much of the forecasting power comes from simply using a different time dummy effect forecast gives us pause in attributing too much of the success to macroeconomic models. Finally, we subject our pooled forecast specification to a robustness check with respect to alternative forecast windows and conclude that even our preferred forecasting procedure cannot consistently outperform the driftless random walk over different forecast windows.

1. Definition of a "Good" Exchange Rate Forecast

There are various criteria for identifying a "good" forecast.[8] One of the most widely used measures, popularized in the exchange rate literature by Meese and Rogoff (1983a), is the

[8]We use the terms "forecast" and "out-of-sample forecast" interchangeably. In order for a forecast to be an out-of-sample forecast, a forecast in period t needs to be a function only of information available in period $t-k$ where the k is the forecast horizon. (For example, if $k=1$ then we are forecasting one period ahead.)

When evaluating the performance of a structural model out-of-sample, we need to be able to compare the forecast produced by the model to the actual realized value of the series we want to forecast. As a result, we split the sample in two – in-sample portion and out-of-sample portion. We run a regression using the in-sample portion and calculate a forecast using the parameters from this regression. We can calculate the forecasts using a recursive or a rolling specification.

minimum mean-square forecast error (MSFE) approach, also known as the MSFE – dominance approach.[9] The goal of this approach is to obtain a model whose MSFE is significantly smaller than that of the random walk model. As Clements and Hendry (1999) suggest, minimum MSFE has become the standard measure of forecast accuracy due to its intuitive interpretation and broad applicability (pp. 9). Another more stringent criterion, introduced by Chong and Hendry (1986), Clements and Hendry (1993), and Harvey et al. (1998) is MSFE encompassing for nested models, which tests whether the structural model encompasses the random walk model. If it does not, then the information provided by the additional explanatory variables does not improve the forecast. MSFE encompassing is more stringent than MSFE dominance, since the latter is a necessary but not sufficient condition for the former. MSFE encompassing also ensures that pooling the competing forecasts cannot produce a forecast with a smaller MSFE than the two nested models considered. A third criterion, robustness over different forecast windows, measures how consistently the structural model outperforms the random walk during different periods of time.

In what follows, we focus first on the minimum MSFE criterion, and afterwards look at robustness over different forecast windows.[10]

2. Minimum Mean-Square Forecast Error Tests: Theil's U (TU), Diebold–Mariano/West (DMW), Clark – West (CW), Clark – McCracken (ENC-NEW)

Before we address the performance of the structural models, we need to revisit the most widely used test statistics in the literature. Until Clark and McCracken (2001, 2005) and Clark and

The recursive method adds one more observation to the in-sample portion for each additional period forecast. For example, if the first forecast is based on the first R observations, then the second forecast is based on the first R+1 observations, etc. In contrast, the rolling specification method preserves the original sample size throughout; hence, the first forecast is based on observations from 1 to R, the second on observations from 2 to R+1, and so on.

[9] Another less popular technique, which our paper does not address, uses the "direction of change" criterion. This criterion, of course, can end up selecting a model which performs well in predicting small changes but poorly at predicting major ones.

[10] We choose not to consider the encompassing criterion for a number of reasons. First, forecast encompassing, defined as the structural model encompassing the random walk, is not widely used in the exchange rate forecasting literature. Second, it is considered a more stringent criterion than MSFE dominance. Third, as Marcellino (2000) points out, the standard encompassing tests may not imply MSFE dominance in the presence of forecast bias. This point is somewhat related to our theoretical argument that the Clark-West and Clark-McCracken out-of-sample tests cannot be always interpreted as minimum MSFE tests in the presence of forecast bias (See Appendix and Section 2 for details).

West (2006, 2007) introduced their tests for nested models, the Theil's U and the Diebold-Mariano/West test statistics were the preferred minimum MSFE out-of-sample test statistics used in the exchange rate forecasting literature. In this paper, we consider the bootstrapped[11] version of both the new and old out-of-sample test statistics (DMW, TU, CW and ENC-NEW) and the asymptotic version of the CW. (For a detailed description of how we calculate each test statistic and how we test statistical significance see Appendix.)

Among the asymptotic test statistics, we focus only on the CW because it has become one of the most popular out-of-sample test statistics for nested models.[12] Furthermore, as we point in the Appendix, the asymptotic versions of the DMW, TU and ENC-NEW have significant shortcomings or are non-tractable. One of the main reasons for the popularity of the asymptotic CW is that the alternative – the use of a bootstrap – is still considered by some researchers computationally cumbersome and difficult to implement.

In this paper we argue that while using the asymptotic CW might seem appealing due to its straightforward application, it is important that one checks the robustness of the results using either the bootstrapped DMW or the bootstrapped TU. The rationale follows.

[11] All of the empirical results presented in the following sections are based on a bootstrap similar to the one used by Mark and Sul (2001). The main difference between our bootstrap and Mark and Sul's (2001) bootstrap is that we use a "semi-parametric" while they use a "parametric" bootstrap and we estimate the error-correction equations using country-specific OLS-regressions rather than seemingly unrelated regressions (SURs) (Note that the "semi-parametric" bootstrap we use is closer in its nature to the "parametric" rather than the "non-parametric" bootstrap. For details on the bootstrap see Appendix).

We choose to use a "semi-parametric" rather than "non-parametric" bootstrap as our preferred bootstrap for a number of reasons. First, based on simulations, Berkowitz and Kilian (1996) argue in their paper "Recent Developments in Bootstrapping Time Series" that when bootstrapping time series, the "parametric" and "semi-parametric" bootstrap outperforms "non-parametric" bootstrap procedures.

Second, the exchange rate forecasting literature provides prolific evidence of the importance of preserving the cointegration between the fundamental and the exchange rate when estimating the exchange rate forecast equation (for example see Kilian, 1999, and Mark and Sul, 2001). And as Berkowitz and Kilian (1996) point out

"While nonparametric bootstrap methods can easily deal with I(1) processes, there are no theoretical results to show that nonparametric resampling preserves cointegration relationships in the data. In fact, cointegration itself may be viewed as a parametric notion. Thus, if the data are known to be cointegrated, parametric methods are preferable (pp. 28)."

For further discussion of cointegration and bootstrapping see Li and Maddala (1997) and Maddala and Kim (1998, pp. 333-336). For completeness sake, we try a number of non-parametric bootstraps such as the wild bootstrap and the block bootstrap but, not surprisingly, their performance is fairly weak and obvious mis-specification problems are present.

[12] The list of studies which test statistical significance using the asymptotic CW includes Engel, Mark and West (2007), Gourinchas and Rey (2007), Molodtsova and Papell (2008), Molodtsova, Nikolsko-Rzhevskyy and Papell (2007, 2008), Rapach, Strauss and Wohar (2007), Sellin (2006), Alquist and Chinn (2006), Cerra and Saxena (2008), Alessi et al. (2007), Groen (2007), Giacomini and Rossi (2008), Ardic et al. (2008), Matheson (2006).

I.) CW/ENC-NEW – Not Always Minimum MSFE Tests One of the main problems related to using the new tests for nested models (CW and ENC-NEW) as the *main and only* out-of-sample test statistics relates to the fact that they cannot be always interpreted as minimum MSFE tests such as the TU and the DMW. In the Appendix we prove that in the presence of forecast bias[13] the CW/ENC-NEW and the DMW do not necessarily test the same null hypothesis; the CW and ENC-NEW test whether the exchange rate is a random walk, whereas TU and DMW test whether the random walk model and the structural model have equal MSFEs. These questions are not equivalent; if the true model is something other than a random walk, one can still perfectly well ask if the random walk model produces a lower mean-square forecast error.[14] However, a significant CW/ENC-NEW and an insignificant bootstrapped TU/DMW can still provide potentially useful information as we show in sections 5 and 6. It implies that, *in theory*, one can pool the forecasts of the structural model and the random walk to produce a combined forecast that outperforms the random walk in terms of MSFE (See Appendix for proof). However, finding an endogenous way of determining this optimal weight has proven to be a challenge (See section 6 for further discussion).

II.) The Asymptotics of CW Are Well Defined Only in the Rolling Case Another problem related to the popular Clark-West out-of-sample test statistic is that the asymptotics of CW are well-defined only when we use the test statistic in a rolling framework, where the size of the in-sample portion of the series is kept fixed. For the recursive case (which comprises the majority of exchange rate forecast specifications in the literature), where the in-sample size varies[15], one has to use simulated critical values based on Brownian motion approximation of the limiting distribution of the CW test statistic.[16] Throughout the paper, the term "asymptotic CW" refers to both the rolling and the recursive case. However, one should keep in mind that in the recursive case the asymptotic distribution of CW is approximated.

III.) Bootstrapped Tests Are Relatively Better Sized and More Powerful Finally, assuming that the bootstrap has been specified correctly, in most specifications, the bootstrapped

[13] Note that by forecast bias we imply only "scale" bias (see footnote 3 for details).

[14] If one tests the explanatory power of the structural model *in-sample* using an ordinary least square (OLS) regression, testing whether the exchange rate is a random walk (testing whether the coefficient in front of the structural model fundamental, b, equals zero) is equivalent to testing whether the random walk has mean square error (MSE) smaller than the MSE of the structural model because OLS minimizes the MSE. However, as the proof of Proposition 1 in the Appendix shows, in *the out-of-sample* case, due to potential forecast bias resulting from forecast uncertainty, testing whether b equals zero is not the same as testing whether the MSFE of the random walk is smaller than the MSFE of the structural model.

[15] See footnote 8 for more details on the distinction between the recursive and rolling specification.

[16] As the authors emphasize, no formal proof is presented that the critical values suggested are appropriate for all forecast specifications (Clark and West, 2007, pp. 298).

DMW and TU out-of-sample tests are more powerful and better sized than the asymptotic CW.[17] Moreover, new research on time series bootstrapping (see for example Li and Maddala, 1997, Berkowitz and Kilian,1996, Kilian, 1999, and Mark and Sul, 2001) and significant improvements in computational power have made the bootstrap an attractive alternative to asymptotic inference.[18]

As a result, we treat the bootstrapped DMW and TU in this paper as our preferred out-of-sample test statistics. In what follows, first we test the robustness of the results of the most popular exchange rate forecasting models and specifications with respect to alternative out-of-sample tests. Second, we concentrate on the robustness of these same specifications with respect to using different sub-samples.

3. Robustness With Respect to Alternative Test Statistics

In this section, we evaluate the robustness of the best results of Gourinchas and Rey (2007), Molodtsova and Papell (2008) and Engel, Mark and West (2007) with respect to alternative test

[17] See the "Not for Publication Appendices" of Clark and West (2006, 2007) that can be found on Kenneth West's website. (Note that in the 2006 Appendix both DGP 1 and 2 are relevant for exchange rate forecasting while in the 2007 Appendix only DGP 1 is of interest.) Regarding comparison between the bootstrapped TU and DMW, see Clark and McCracken (2005).

The concepts of size and power are key to understanding the differences between the alternative out-of-sample test statistics. They are properties of both the asymptotic and bootstrapped tests. The size of a test statistic is defined as the test's probability of rejecting the null hypothesis if the null is true. If the researcher chooses to use a significance level of 10%, an under-sized(over-sized) test statistic would tend to reject the null hypothesis in less(more) than 10% of the cases. If a test statistic is over-sized, it might incorrectly detect statistical significance if such does not exist and if it is under-sized – incorrectly reject the alternative. The power of a test statistic is defined as the test's probability of correctly rejecting the null hypothesis for a given level of statistical significance. The size and power of a test statistic are inversely related.

In the Appendix of Clark and West (2006), it is not immediately obvious why the bootstrapped DMW has greater power than the asymptotic CW because the authors report size-adjusted power rather than raw power. The main difference between the two is that only raw power is of any practical importance since in order to adjust for size distortions, the size-adjusted power is based on a CW test statistic which uses data specific critical values obtained via Monte Carlo simulation. Since few, if any, researchers would choose this alternative, the raw power is what one is mainly interested in. Given that the size-adjusted power of CW is similar to that of the bootstrapped DMW, the raw power of CW will be smaller than the raw power of the bootstrapped DMW. This is the case because the CW is somewhat undersized while the bootstrapped DMW seems adequately sized and as we already explained the size and power of a test statistic are inversely related.

Finally, according to the simulation evidence in Clark and McCracken (2005), we would expect the bootstrapped TU to be more powerful than the bootstrapped DMW. (Note that in their paper the authors discuss the power of the MSE-F rather than the TU but the two tests are very similar). Since the bootstrapped DWM is more powerful than the asymptotic CW, we would expect that the bootstrapped TU is more powerful than the asymptotic CW as well.

[18] See footnotes 6 and 11 and Appendix for further discussion on time series bootstrapping.

statistics (bootstrapped CW, ENC-NEW, DMW and TU). These studies all feature the asymptotic CW as their main out-of-sample test statistic and conclude that for a number of countries, structural models outperform the driftless random walk for forecasts one period ahead. While Engel, Mark and West (2007) attribute their success to the power of panel models, Molodtsova and Papell (2008) and Gourinchas and Rey (2007) find success using new structural models. While we concentrate our attention on these three prominent studies with fairly positive results, we believe that the implications of our findings are generalizable to the rest of the new literature on short-horizon exchange rate forecasting.

We conclude that forecast bias is a serious problem which in certain specifications leads to a significant discrepancy between the CW/ENC-NEW and DMW/TU. Furthermore, in a couple of cases, the asymptotic CW is oversized. As a result of both of these issues, some of the results of the literature are overly optimistic and potentially misleading.

Engel, Mark and West (2007) - The Monetary Model

The implementation of a panel forecast specification is one of the key additions to the exchange rate forecasting literature which allows Engel, Mark and West (2007) to find limited forecastability of the exchange rate change one quarter ahead.[19] The study finds that for 5 out of 18 currencies, the monetary model outperforms the driftless random walk. While recognizing this success as modest, the authors note that their results appear notably more positive than the norm in the literature.[20]

The forecasting specification Engel, Mark and West (2007) apply is straightforward. The forecast variable is the nominal exchange rate change, where s_t is the natural log of the exchange rate measured in foreign currency per one unit of the base currency (in this case US dollars). Define $\Delta s_{i,t+1} = s_{i,t+1} - s_{i,t}$ and the forecast is one period ahead. Then the panel forecast equation can be expressed as

$$\Delta s_{i,t+1} = \alpha_i + \theta_t + \beta z_{i,t} + \varepsilon_{i,t+1}. \tag{1}$$

where, in this case, $z_{i,t}$ stands for the deviation of the exchange rate from an equilibrium value. $z_{i,t}$ is determined by the monetary model fundamental

[19] The study of Engel, Mark and West (2007) builds on Engel and West (2005).

[20] The majority of the recent panel specification papers find strong support for the forecasting power of the monetary model in both long and short horizons (see Mark and Sul, 2001, Rapach and Wohar, 2002, Engel, Mark and West, 2007, and Groen, 2005). However, at the same time, the theoretical validity of the monetary specification has been widely criticized. The criticism of the monetary model centers around its assumptions that both purchasing power parity and uncovered interest parity hold. However, these assumptions are not unequivocally supported by empirical evidence (Engel, 1996). Furthermore, there is a debate on how one defines the money supply, the stability of the money equation (Friedman and Kuttner, 1992) and whether money has any relevance for economic decision making such as monetary policy.

$$z_{i,t} = m_{i,t} - m_t^* - \varphi(y_{i,t} - y_t^*) - s_{i,t}. \quad (2)$$

Above, i is a country-specific index, α_i stands for country-specific effects, θ_t - for time specific effects and $\varepsilon_{i,t+1}$ is the innovation term. The $(*)$ represents the base country (the US), $(m_{i,t} - m_t^*)$ is the relative money supply, $(y_{i,t} - y_t^*)$ is the relative income level and φ is assumed to be one. Note that all the variables are in natural logs. We use the driftless random walk, expressed as $\Delta s_{i,t+1} = \upsilon_{i,t+1}$ (where $\upsilon_{i,t+1}$ is the innovation term of the driftless random walk model) as a benchmark which would ensure that the structural model is compared to the best known alternative.[21]

Engel, Mark and West (2007) estimate equation (1) using recursive OLS regressions. They calculate the exchange rate change forecast using the following equations.

$$\text{Structural Model} : \Delta\hat{s}_{i,t+1} = \hat{\alpha}_i + \hat{\theta}_{t+1} + \hat{\beta} z_{i,t+1}$$

$$\text{Driftless Random Walk Model} : \ \Delta\hat{s}_{i,t+1} = 0$$

where the time dummy for period $t+1$ is calculated as $\hat{\theta}_{t+1} = \frac{1}{t}\sum_{j=1}^{t}\hat{\theta}_j$. Engel, Mark and West's (2007) sample extends Mark and Sul's (2001) data set up to 2005Q4. The exchange rates of the Eurozone countries post 1999 are normalized in a way that they differ from each other only by a constant.[22] This implies that post 1999, Engel, Mark and West's (2007) specification is essentially forecasting the same exchange rate - the Euro - using different country specific monetary fundamentals. For further details on the specification and for data set sources refer to Engel, Mark and West (2007).[23]

We test the robustness of their results with respect to different test statistics. In Table 1, we reproduce the monetary model results but rather than just report the asymptotic CW test statistic, we also report the bootstrapped p-values of the DMW, TU, CW and ENC-NEW. If we test statistical significance via the bootstrapped DMW and TU test statistics, the p-value is less than 10% for only 4 out of 18 cases.[24] These results are confirmed by the bootstrapped

[21] Engel, Mark and West (2007) compare the forecasts of the monetary model to both the random walk with drift and without drift. However, they note that the driftless random walk outperforms the random walk with drift. All of the studies we are aware of that compare the driftless random walk to the random walk with drift, find the driftless random walk to be a better forecaster (see Engel and Hamilton, 1990, and Engel, Mark and West, 2007).

[22] For example, the normalization for France post 1999 will be simply franc/euro times euro/dollar where the franc/euro is the peg used to fix the French franc to the euro in 1999.

[23] We use Engel, Mark and West's (2007) data except for exchange rates which are from the IFS data set. The bootstrap procedure is similar to Mark and Sul (2001) and assume no unit root of the monetary fundamental. (For details on the bootstrap used see Appendix.)

[24] Another way of testing for robustness, which we do not pursue in this paper, is by estimating to what extent the positive results could be attributed to the large number of specifications tested. For instance, the test statistic introduced by McCracken and Sapp (2005) tests whether the number of successful forecasts can be attributed solely to the large number of specifications and models estimated by the researcher. If we were to calculate McCracken and Sapp's (2005) test statistic, the results might have been even less favorable for the structural models.

CW. However, Greece stands out as an example where the asymptotic CW is statistically significant while the bootstrapped CW is not, suggesting that the asymptotic CW is over-sized.[25]

Table 1: Mark, Engel and West (2007)
Monetary Model Vs Random Walk with No Drift; One Quarter Ahead

		Reproduced Results							
	CW^	**CW**	**P-value**	**TU**	**P-value**	**DMW**	**P-value**	**ENC-NEW**	**P-value**
UK	0.684	0.624	0.25	1.001	0.262	-0.063	0.254	0.739	0.251
Austria	**1.966**	**1.854**	**0.04**	**0.984**	**0.040**	1.315	**0.015**	2.164	0.110
Belgium	1.199	0.972	0.13	1.001	0.327	-0.041	0.292	1.653	**0.034**
Denmark	0.259	0.178	0.372	1.009	0.657	-0.553	0.459	0.257	0.340
France	0.706	0.545	0.23	1.001	0.228	-0.050	0.225	0.543	0.237
Germany	**1.855**	**1.711**	**0.056**	**0.986**	**0.048**	0.924	**0.065**	2.408	**0.080**
Netherlands	**1.651**	**1.400**	**0.084**	**0.990**	**0.071**	0.922	**0.060**	1.375	0.144
Canada	-0.942	-0.936	0.881	1.161	0.970	-2.348	0.906	-4.240	0.988
Japan	1.094	0.671	0.439	**0.999**	0.366	0.038	0.364	0.873	0.425
Finland	0.648	0.696	0.262	1.004	0.420	-0.156	0.339	1.463	0.154
Greece	**2.509**	**2.501**	0.704	1.004	0.906	-0.085	0.899	11.211	0.450
Spain	0.711	0.699	0.592	1.027	0.916	-0.806	0.799	2.091	0.343
Australia	0.787	0.727	0.343	1.026	0.555	-0.914	0.479	1.869	0.303
Italy	0.733	0.557	0.519	1.015	0.764	-0.487	0.575	1.525	0.401
Switzerland	**1.965**	**1.985**	**0.06**	**0.986**	**0.082**	1.409	**0.019**	1.912	0.197
Korea	0.853	0.847	0.385	**0.997**	0.252	0.118	0.306	1.972	0.188
Norway	0.645	0.271	0.327	1.005	0.426	-0.402	0.358	0.296	0.317
Sweden	1.100	1.030	0.225	**0.993**	0.163	0.466	0.174	1.372	0.212

^Results provided by Charles Engel using a corrected data set

The benchmark is a random walk without drift; Quarterly data ranging from 1973Q1 to 2005Q4; First Forecast: 1983Q1; The p-value is the bootstrapped version of the respective test statistic. Bootstrap based on 1000 iterations; Bold p-values imply statistical significance of 10% or less; Bold Theil's U values represent Theil's U <= 1; Bold CW values represent statistical significance of 10% (above 1.282) using Clark and West's (2007) simulated critical values.

[25] In the case of Greece, the asymptotic CW performs so poorly because the DMW is largely oversized and, as a result, the asymptotic CW is even more over-sized (see Appendix for clarification on the difference between CW and DMW). The mean of the bootstrapped DMW histogram is 1.3 (and it should have been 0 if no size problem was present). As is apparent from the results in Table 1, in outlier cases like Greece, the asymptotics fail while the bootstrap, if properly specified, seems to be still reliable.

Molodtsova and Papell (2008) -Heterogeneous Symmetric Taylor Rule with Smoothing

In addition to the improvements produced by the panel specification, the introduction of the Taylor rule as a structural fundamental has also seemed to yield improved forecasts. The specification which produces best forecasting results estimates country specific coefficients on both inflation and the output gap. Furthermore, Molodtsova and Papell (2008) assume that interest rates adjust only partially to its target and, as a result, lagged interest rates are included in the specification which represent the so-called smoothing effect. Using only single-country equations and the asymptotic CW test statistic, Molodtsova and Papell (2008) conclude that the Taylor rule outperforms the driftless random walk for 10 out of 12 currencies for forecasts one period ahead. Molodtsova and Papell (2008) specify the fundamental, z_t, as

$$z_t = \alpha_1 \pi_t + \alpha_2 \pi_t^* + \alpha_3 y_t^{gap} + \alpha_4 y_t^{gap*} + \alpha_5 i_{t-1} + \alpha_6 i_{t-1}^* \tag{3}$$

where π is the inflation rate, i is the interest rate and y^{gap} is the output gap defined as the deviation of an industrial production index from a linear trend. We substitute equation (3) in equation (1) and estimate equation (1) in a single equation framework using monthly data.[26] Molodtsova and Papell (2008) refer to specification (3) as the heterogeneous symmetric Taylor rule with smoothing.[27] More information regarding the specification and data sources is provided in Molodtsova and Papell (2008).

Similarly to the monetary model specification of Mark, Engel and West (2007), we replicate Molodtsova and Papell's (2008) results and compute not only the asymptotic CW, but also the bootstrapped TU, DMW, CW and ENC-NEW.[28] Table 2 reports Molodtsova and Papell's (2008) results as presented in their paper and our attempt to replicate them using their methodology and data set.

There is a striking difference between both the bootstrapped and asymptotic CW and the bootstrapped ENC-NEW, on the one hand, and the bootstrapped TU and DMW on the other hand. While CW and ENC-NEW are significant in as many as 10 out of 12 cases, the TU is not significant for any of the countries and DMW is significant only for Canada. We explain this discrepancy with the presence of severe forecast bias in which case the CW and ENC-NEW cannot be interpreted as minimum MSFE tests and they do not test the same null hypothesis

[26] Single-equation framework implies that there are no time dummy effects.

[27] The estimation method is a rolling regression specification with a rolling window of 120 months.

[28] The data set was provided to us by the authors and it is also available on David Papell's website. For the bootstrap, similarly to the bootstrap used to replicate the results of Engel, Mark and West's (2007) study, we use similar to Mark and Sul's (2001) procedure. We assume that the inflation rates, the interest rates and the output gaps do not have unit roots. (For details on the bootstrap used see Appendix.)

as TU and DMW.[29] In the case of Switzerland, the bootstrapped CW is insignificant while the asymptotic CW is significant, again, suggesting that the asymptotic CW might be oversized in certain cases.

Table 2: Molodtsova and Papell (2008)
Heterogeneous Symmetric Taylor Rule with Smoothing Vs Random Walk with No Drift; One Month Ahead

	CW P-value asymptotic^	CW P-value asymptotic	CW P-value bootstrap	TU	P-value	DMW	P-value	ENC-NEW	P-value
UK	**0.020**	**0.027**	**0.027**	1.051	1.000	-1.740	0.678	14.662	**0.001**
Denmark	**0.069**	**0.067**	**0.045**	1.025	0.992	-1.231	0.397	8.067	**0.013**
France	**0.024**	**0.019**	**0.007**	1.040	0.998	-1.260	0.557	11.312	**0.001**
Germany	**0.066**	**0.066**	**0.077**	1.036	0.997	-1.130	0.548	8.458	**0.016**
Netherlands	**0.036**	**0.035**	**0.040**	1.040	1.000	-1.304	0.613	9.604	**0.012**
Canada	**0.008**	**0.008**	**0.008**	1.006	0.174	-0.261	**0.078**	15.025	**0.003**
Japan	**0.019**	**0.019**	**0.071**	1.018	0.912	-0.723	0.367	14.152	**0.008**
Australia	**0.015**	**0.013**	**0.039**	1.024	0.972	-0.895	0.360	15.130	**0.004**
Italy	**0.002**	**0.002**	**0.039**	**0.995**	0.264	0.168	0.327	18.240	**0.003**
Switzerland	**0.094**	**0.094**	0.153	1.068	1.000	-2.198	0.910	9.151	**0.021**
Sweden	0.678	0.674	0.667	1.098	1.000	-1.261	0.494	-5.897	1.000
Portugal	0.985	0.985	0.985	1.127	1.000	-3.329	0.999	-4.464	1.000

^ Results as reported in Molodtsova and Papell (2008)

Single equation, monthly data. Since Molodtsova and Papell (2008) use rolling regressions, the asymptotic CW p-values are calculated under the assuming of normality; The TU, ENC-NEW and DMW p-values and the CW bootstrap p-value are based on a bootstrap (1000 iterations); Bold Theil's U values represent Theil's U <= 1;

Gourinchas and Rey (2007) - External Balance Model

Another important study that claims to successfully forecast exchange rates one period ahead is Gourinchas and Rey (2007). The authors introduce a new external balance model which isolates long- term effects by defining an external balance variable as a function of de-trended foreign assets and liabilities, exports and imports. Gourinchas and Rey (2007) find that their external balance measure is superior to those previously used in the literature on external balance specifications since it takes into account capital gains and losses on the net foreign asset position,

[29] A regression of the observed exchange rate change on the forecast series and no constant produces a coefficient less than or close to 0.5 for all 10 countries where CW and ENC-NEW are significant. (If no "scale" forecast bias was present, the coefficient should have been close to 1.) This is what we would expect in cases of severe "scale" forecast bias which can lead to CW and ENC-NEW not testing the same null as TU and DMW.

in addition to the trade balance. Gourinchas and Rey (2007) argue that their external balance variable successfully forecasts both the trade and FDI-weighted dollar one quarter ahead.

We can write Gourinchas and Rey's (2007) external balance fundamental as

$$z_t = |\mu_t^a|\epsilon_t^a - |\mu_t^l|\epsilon_t^l + |\mu_t^x|\epsilon_t^x - |\mu_t^m|\epsilon_t^m \tag{4}$$

where $\mu_t^a, \mu_t^l, \mu_t^x$ and μ_t^m are time-varying weights – a function of the Hodrick-Prescott-filtered trends of assets, liabilities, exports and imports – while $\epsilon_t^a, \epsilon_t^l, \epsilon_t^x$ and ϵ_t^m represent the log deviation of assets, liabilities, exports and imports from Hodrick-Prescott-filtered trends. Equation (4) is substituted into equation (1) and the authors estimate equation (1) for the trade-weighted and the FDI-weighted exchange rate separately in a single equation framework.[30] Gourinchas and Rey (2007) assume that the time-varying weights converge asymptotically and use fixed weights for the calculation of their forecasts. Further details on the specification and the data set used are provided in Gourinchas and Rey (2007). In Table 3, we reproduce their results using their data set and similar methodology.[31] One can observe highly significant asymptotic CW, bootstrapped TU, DMW, CW and ENC-NEW test statistics. However, the seemingly strong result is overturned, to an extent, when checking for robustness with respect to alternative time periods in the following section.

[30]Note that the authors claim to be using a 105 quarter rolling window. However, a closer look at their code shows that they use 105 quarter rolling window for the forecasts of the FDI-traded dollar and a recursive specification for the trade-weighted dollar. We calculate the forecast both ways – in a recursive and rolling framework – and the results do not change substantially.

[31]We are grateful to the authors for providing us with their code and data set. Note that in Table 3 we report the CW test statistic which is calculated as $CW = \frac{P^{0.5}\hat{d}}{\sqrt{\Omega^{\hat{d}}}}$ where $\hat{d}$ is defined in equation (6) in Appendix, while Gorinchas and Rey (2007) report $\hat{d}$ in their paper.

As before, we use a bootstrap procedure similar to Mark and Sul (2001) and assume no unit root of the external balance variable. (For details on the bootstrap used see Appendix.)

Table 3: Gourinchas and Rey (2007)
External Balance Model Vs Random Walk with No Drift; One Quarter Ahead

	Reported by G&R	Reproduced Results							
	CW	CW	P-value	TU	P-value	DMW	P-value	ENC-NEW	P-value
Trade Weighted Exch Rate	**2.690**	**2.684**	**0.005**	0.974	**0.003**	0.657	**0.013**	**11.774**	**0.001**
FDI Weighted Exch Rate	**2.196**	**2.191**	**0.006**	0.980	**0.005**	0.821	**0.019**	**5.780**	**0.002**

FDI - Weighted exch rate - rolling regression where the rolling window is 105 quarters; first forecast: 1978 Q3
Trade - Weighted exch rate - recursive regression; first forecast: 1978 Q2

The p-value is the bootstrapped version of the respective test statistic. Bootstrap based on 1000 iterations; Bold p-values imply statistical significance of 10% or less; Bold Theil's U values represent Theil's U <= 1; Bold CW values represent statistical significance of 10% (above 1.282) using Clark and West's (2007) simulated critical values.

Summary of Test Statistic Robustness

We looked at each one of the three major studies which find forecastability one period ahead and concluded that when one considers the robustness of the results with respect to alternative test statistics, the results of Molodtsova and Papell (2008) fluctuate significantly due to the presence of forecast bias. The results of Mark, Engel and West (2007) are somewhat less spectacular as a result of one outlier where the asymptotic CW is severely oversized. Finally, we conclude that the results of Gourinchas and Rey (2008) remain robust to the test statistic considered. Now we turn to our second main issue – which of the results are robust over different periods of time.

4. Robustness with Respect to Different Forecast Windows

In addition to testing the robustness of the results with respect to different out-of-sample test statistics, we also test the robustness of the results of Molodtsova and Papell (2008), Engel, Mark and West (2007) and Gourinchas and Rey (2007) via varying the forecast window. This is another important test of how consistently reliable the forecast is.[32]

We find that the structural models do not produce consistently better forecasts than the driftless random walk over different sample periods for the three studies reviewed. However, it seems that the monetary model and the Taylor rule model forecast the exchange rate better than

[32]Giacomini and Rossi (2008) is one of the few studies which attempts to formalize the issue of robustness over different forecast windows by developing appropriate test statistics.

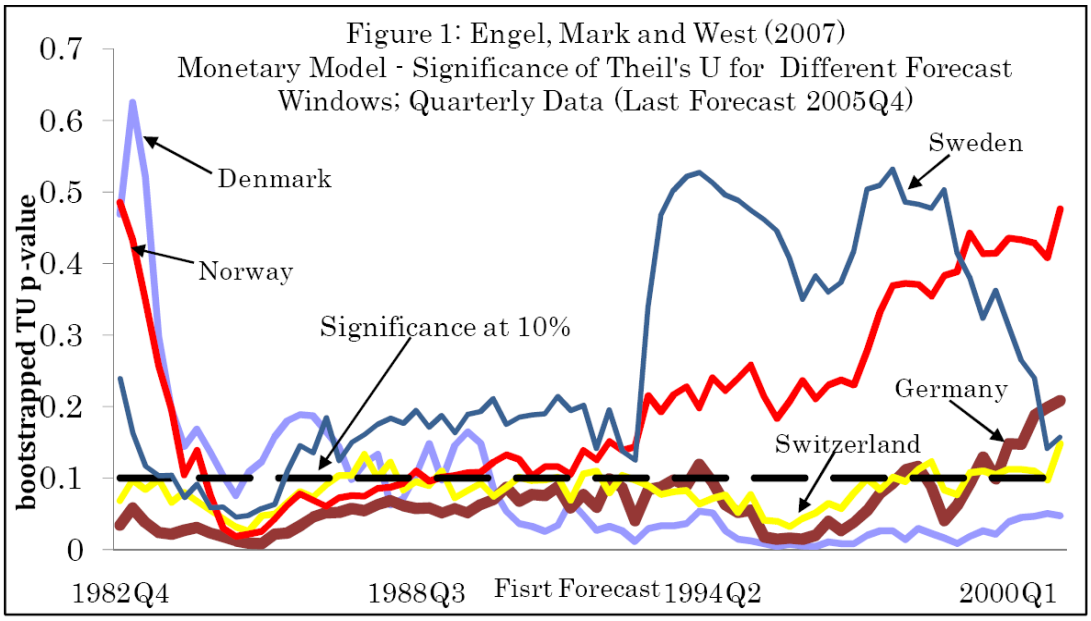
the random walk during the 1980s while Gourinchas and Rey's (2007) external balance model consistently outperforms the driftless random walk in the 1990s and the 2000s.

Engel, Mark and West (2007)

Figure 1 illustrates an approach to testing robustness to different sample periods in the context of Engel, Mark and West's (2007) monetary model. Figure 1 plots the bootstrapped TU (Theil's U) p-value on the y-axis and the starting date of the recursion on the x-axis (the first date for which a forecast is calculated). We plot only the bilateral exchange rates for which we find forecastability for a large number of forecast windows in order to make the graph legible. For similar reasons, we report only the results for Germany as a proxy for the Eurozone countries.

The way Figure 1 should be interpreted is the following. For example, the TU p-value associated with 1984Q4 for a given country implies that the TU p-value is calculated using the forecast window from 1984Q4 to 2005Q4 (the end of the sample). If the TU p-value is below 0.1, we consider the result statistically significant at 10 percent. In order for a result to be considered robust, we would expect that the TU p-value is below 0.1 for almost all of the plotted forecast windows. The graph shows that the monetary model is a relatively good forecaster of the Swiss franc and, to a lesser extent, of the Deutsche mark/euro.[33] It is interesting to note that overall (when one considers the other current Eurozone countries as well), the monetary model performs relatively well in the 1980s and its performance deteriorates in the 1990s.

[33] Keeping in mind that post-1999 the Deutsche mark transitions into the Euro, the fact that the TU test statistic for Germany becomes insignificant when we restrict the forecast window post year 2000 implies that while the monetary model was a relatively good forecaster of the Deutsche mark, this might not be the case for the Euro. However, with more euro data the result could change.



In conclusion, while at first look (considering one forecast window only), Engel, Mark and West's (2007) results seem encouraging, if one considers the robustness of the results over different forecast windows, they are less so.[34]

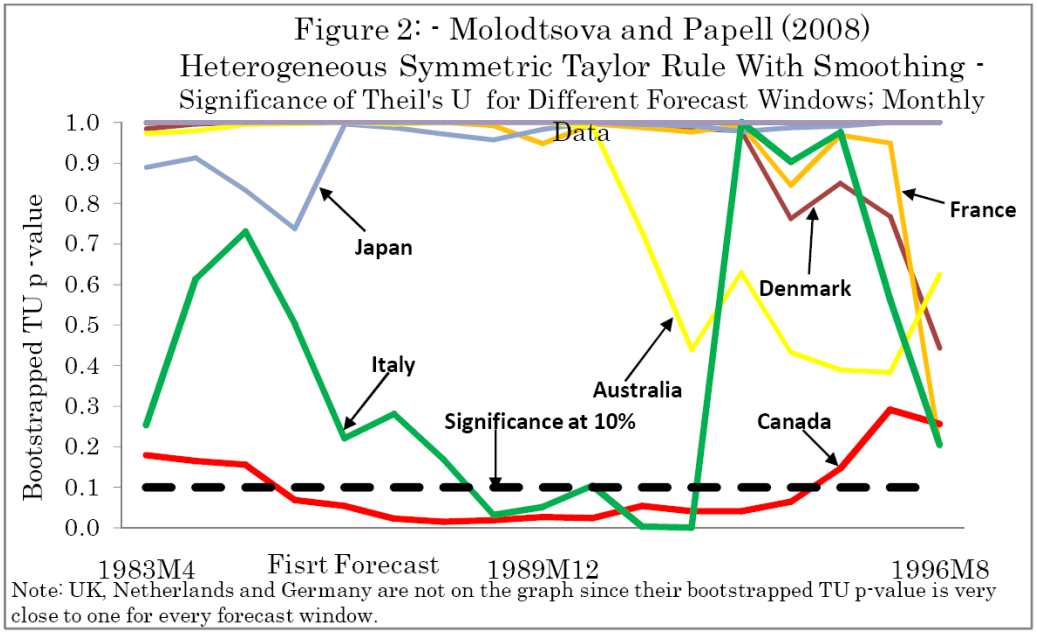
Molodtsova and Papell (2008)

In a similar fashion, we evaluate Molodtsova and Papell's (2008) heterogeneous symmetric Taylor rule results with smoothing for consistency over different forecast windows. Figure 2 depicts the robustness of Molodtsova and Papell's (2008) results with respect to starting the rolling regression at a different date. Only the countries for which the bootstrapped CW are significant are reported (nine out of twelve). Figure 2 clearly shows that the results are somewhat robust only for Canada – the only country for which the TU is below 0.1 for most of the forecast windows.[35]

[34]The results remain non-robust when one reproduces Figure 1 plotting the bootstrapped ENC-NEW/CW test statistic rather the bootstrapped TU p-value.

[35]While interesting, the result for Canada is perhaps not that surprising given that Reinhart and Rogoff (2002) classify Canada as a limited flexibility exchange rate.

Even though, in Molodtsova and Papell's (2008) specification, one cannot interpret the CW and ENC-NEW as minimum MSFE tests due to the presence of severe forecast bias, for completeness, we reproduce Figure 2 using the bootstrapped CW. The bootstrapped CW is significant in the early 1980s for the majority of the countries but not significant for the rest of the period (the only exceptions are Australia, Canada, Italy and Japan for which the bootstrapped CW is significant for the majority of forecast windows).



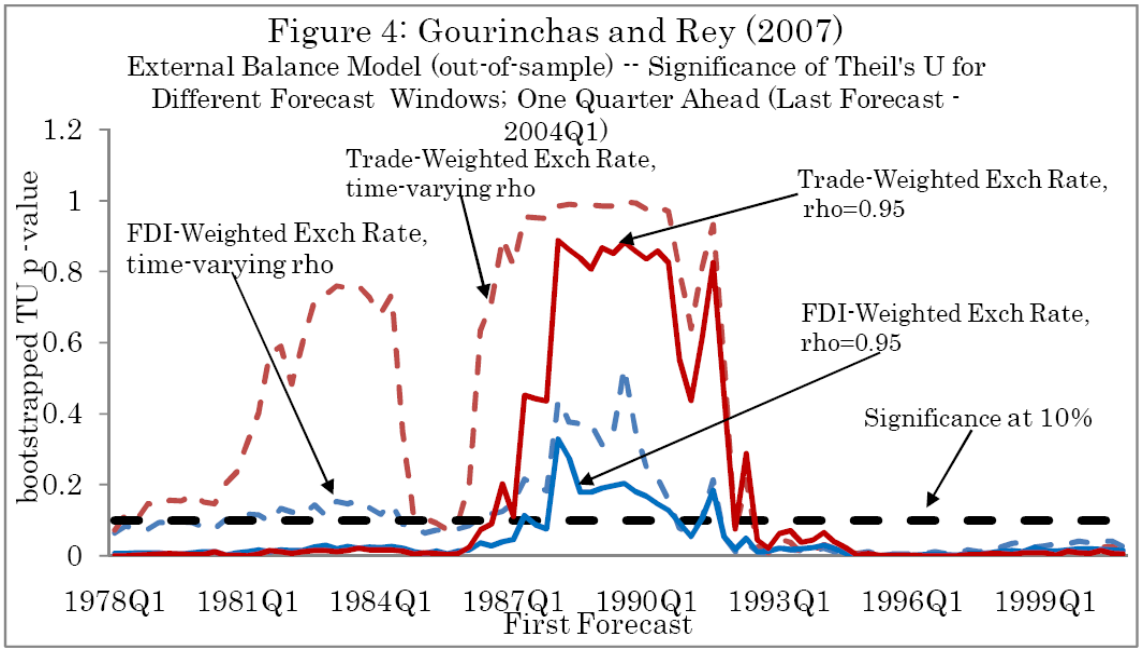
Gourinchas and Rey (2007)

Finally, in order to test to what extent the results of Gourinchas and Rey's (2007) external balance model are robust to changing the forecast window, we report the p-value of the bootstrapped TU test statistic for different forecast windows in Figure 3. The forecasts are estimated using a recursive specification.[36] We report two approaches of estimating the external balance variable. The solid lines in Graph 3 represent Gourinchas and Rey's (2007) approach which estimates the external balance variable as a function of a discount constant, ρ, which is assumed to be 0.95 (a long-term value obtained using the entire sample). In order to test the robustness of Gourinchas and Rey's (2007) results to alternative values of ρ, we calculate a time-varying discount rate (as a opposed to a long-run value) using only the in-sample portion of the data. This approach is presented with dashed blue and red lines in Graph 3.

As before, we consider TU significant at 10 percent when the bootstrapped p-value is below the dashed black line. It is interesting to note that the FDI-weighted exchange rate performs relatively well for most of the periods with the exception of the late 1980s. The performance of the trade-weighted series is less impressive (especially when one considers

[36]If we calculate the forecasts using a rolling window of 105 quarters (the rolling window Gourinchas and Rey, 2007, state they use) rather than recursive regressions, the results do not change substantially. However, using recursive regressions allows us to check the robustness of the trade-weighted dollar for different forecast windows given the shorter range of the series. We also calculate time varying weights rather than impose constant weights which also seems to affect the final results only negligibly.

the time-varying ρ approach which calculates the discount rate using only in-sample information). However, no matter whether one considers the trade-weighted or the FDI-weighted series, the external balance variable outperforms the driftless random walk in the 1990s and 2000s.[37]



Summary of Robustness to Perturbation in Sample

After examining the performance of the monetary model, the external balance model and the heterogenous Taylor rule with smoothing over different periods of time, we conclude that Engel, Mark and West's (2007) monetary model specification performs well in the 1980s but poorly in the more recent period. Molodtsova and Papell's (2008) Taylor rule results, which were not robust to the use of the bootstrapped TU and DMW in section 3, are also highly non-significant for all the periods considered. It seems that the performance of Gourinchas and Rey's (2007) external balance variable is potentially more encouraging for the most recent period. However, one should be cautious in comparing the performance of structural models which forecast bilateral exchange rates with those which forecast weighted exchange rates, as the latter tend to be significantly less volatile.

[37]The results are similar if we consider the rest of the out-of-sample test statistics.

5. Can One Do Better? The Importance of Common Cross-Country Shocks

In this section we try to improve upon the panel forecast specification applied by Mark and Sul (2001) and Engel, Mark and West (2007) by incorporating persistent common cross-country shocks in the forecasts. These might include technology shocks, commodity price shocks, or factors related to the pace of globalization.

The basic forecast specification we use is the same as the one defined in equation (1). However, we define $z_{i,t}$, the deviation of the exchange rate from an equilibrium value, using the purchasing power parity model (PPP) rather than the monetary model.[38]

$$z_{i,t} = p_{i,t} - p_t^* - s_{i,t} \tag{5}$$

Above, p is the natural log of the CPI and, as before, the $(*)$ represents the US. We substitute equation (5) into equation (1) and estimate equation (1) using recursive OLS panel regressions. The way we take into account potential persistent cross-country shocks is by forecasting the time dummy effect for period $t+1$ differently from previous panel studies. Rather than forecast it simply as the average of the time-dummy coefficients for all the previous periods, as Mark, Engel and West (2007) did, we forecast it as a simple average of the last 4 estimated time-dummy coefficients. Mathematically, the time dummy forecast can be defined as

$$\theta_{t+1} = \frac{1}{q} \sum_{j=t-q+1}^{t} \theta_j$$

where $q = 4$ when the data is quarterly.

Table 4 reports the results of the specification defined above. [39]

[38]The PPP specification is known to perform well at long horizons, but has been much less explored in looking at short-horizon nominal exchange rate forecasts. Engel, Mark and West (2007) is the only study, we are aware of, which has explored the forecasting power of the PPP model at short horizons in a panel framework where the benchmark is the driftless random walk. Engel, Mark and West (2007) find that for forecasts one period ahead, the PPP forecast is significantly better than the driftless random walk forecast only in 3 out of 18 cases.

We also perform the same type of forecasting exercise as in sections 5 and 6 using the monetary model, the Taylor rule and a new structural model based on the Backus - Smith optimal risk sharing condition model. Out of all the models we try, the PPP specification performs the best.

[39]The data source is the International Financial Statistics (IMF) (Data available upon request). Our data set consists of eleven countries: US, UK, Denmark, Germany, Canada, Japan, Australia, Switzerland, Korea, Norway and Sweden. We choose to proxy the Euro using the Deutsche mark series up to 1999 and the euro post 1999. The bootstrap specification is similar to Mark and Sul (2001) and the same as the bootstrap used in the literature review section (see Appendix for details).

Table 4: PPP Specification; One Quarter Ahead; Forecasts Incorporate Common Cross-Country Shocks

	CW	P-value		ENC-NEW	P-value		Theil's U	P-value		DMW	P-value
UK	0.423	0.326		1.343	0.242		1.072	0.991		-1.879	0.810
Denmark	**1.469**	**0.072**		3.884	**0.040**		1.008	0.268		-0.285	0.226
Germany	**1.662**	**0.075**		5.655	**0.034**		**0.999**	0.145		0.035	0.147
Canada	**2.724**	**0.004**		21.411	**0.000**		1.056	0.444		-0.665	0.142
Japan	**1.809**	**0.090**		5.592	**0.056**		1.004	0.280		-0.118	0.265
Australia	**2.435**	**0.021**		11.391	**0.003**		**0.958**	**0.006**		0.959	**0.026**
Switzerland	1.239	0.230		4.184	0.106		1.008	0.445		-0.250	0.385
Korea	**1.870**	**0.059**		6.634	**0.015**		**0.975**	**0.007**		0.712	**0.056**
Norway	0.944	0.155		2.363	0.115		1.033	0.741		-1.232	0.534
Sweden	**2.215**	**0.030**		5.398	**0.035**		**0.999**	0.155		0.028	0.157

Recursive specification; quarterly data; country and time dummies included; time dummy effect forecasted as the simple avg of estimated time dummies over last 4 quarters; first forecast 1983Q1; last forecast 2007Q1 (PPP); The p-value is the bootstrapped version of the respective test statistic. Bootstrap based on 1000 iterations; Bold p-values imply statistical significance of 10% or less; Bold Theil's U values represent Theil's U <= 1; Bold CW values represent statistical significance of 10% (above 1.282) using Clark and West's (2007) simulated critical values.

The results of Table 4 are very similar to the results in Molodtsova and Papell (2008) presented in Table 2. If we concentrate our attention only on the statistical significance of the bootstrapped TU and DMW test statistics, we notice that the results are significant only for Australia and Korea. However, when one calculates the CW and ENC-NEW out-of-sample test statistics, CW and ENC-NEW are significant for 7 out of 10 countries. Note that the bootstrapped CW behaves similarly to the asymptotic CW. In combination with the fact that Clark and McCracken (2005) and Clark and West (2006, 2007) conclude that the bootstrapped DMW and TU tend to be more powerful than the asymptotic CW and adequately sized, this implies that the discrepancy between the CW/ENC-NEW and TU/DMW cannot be attributed to different power and size. Furthermore, an investigation of the results indicates the presence of forecast bias.[40] As a result, the only explanation left for the discrepancy is that the two types of test statistics test different null hypotheses and cannot be used interchangeably – a point we prove and discuss further in the Appendix.

6. Forecast Pooling – Empirical Example

[40]Substantial "scale" forecast bias is present in all of the cases where we observe a discrepancy between the TU/DMW and CW/ENC-NEW.

However, a significant CW/ENC-NEW still provides useful information when the bootstrapped TU/DMW is insignificant which is what one observes in Table 4 (For proof see Appendix). In this section we provide an empirical example of how in such cases one can improve upon the structural model forecast by pooling the structural model forecast and the random walk forecast.

Endogenous vs Exogenous Weights

The question emerges how one can calculate a weight which will produce a forecast with MSFE smaller than the MSFE of the random walk. One can either use endogenous time-varying methods of finding the optimal weight (see Clements and Hendry, 1998, pp. 229), or one can impose a fixed weight exogenously. It is conventional wisdom in the literature on forecast pooling that simple averages tend to outperform endogenous weights (See Stock and Watson, 2003, Clark and McCracken, 2006 and Clements and Hendry, 2004). Clements and Hendry (2004) explain this phenomena with the fact that all endogenous procedures of finding an optimal weight would be biased in the presence of structural breaks (which might be one explanation of the lack of robustness of the models over different forecast windows discussed in section 4).[41] In contrast, having a constant weight can serve as an insurance against structural breaks and perform overall better than a time-varying endogenous weight.[42]

We test which pooling procedure produces better results – imposing exogenous fixed weights or calculating endogenous weights using the regression method presented in Clements and Hendry (1998, pp. 229). As expected, our results confirm the conclusion of the literature on forecast pooling that simple means and fixed weights perform better than endogenously calculated optimal weights. [43] As a result, we choose to impose a fixed weight of 0.2 on the structural model forecast and 0.8 on the random walk forecast (which is essentially zero).[44]

[41] Since endogenous weights are estimated on the basis of data prior to the forecast, a structural break in the recent past or in the near future will lead to biased weights. It is possible that prior to the break, a certain model performs better than the alternative but performs poorly after the structural break. As a result, endogenously determined weights would lead to the forecaster weighting more heavily the model which performed better prior to the break but poorly after it.

[42] Potential structural breaks affect also the degree to which the information provided by the CW and the ENC-NEW is valuable. In the presence of structural breaks, pooling can be appropriate even if the CW and the ENC-NEW are not statistically significant (and of course the bootstrapped TU and DMW are not statistically significant) (see Hendry and Clements, 2004). The reason why this is the case is that the forecaster does not know in advance whether the CW/ENC-NEW will be significant or not if the test statistic is calculated using data from the next regime.

[43] Results available upon request.

[44] The results are relatively robust to using a simple average.

Results After Pooling

Table 5 presents the "pooled" forecast results where we combine the forecast of the PPP model which incorporates persistent cross-country shocks and the driftless random walk model.

Table 5: "Pooled" Forecast; PPP Specification; One Quarter Ahead; Forecasts Incorporate Common Cross-Country Shocks

	Theil's U	P-value		DMW	P-value
UK	1.000	0.452		-0.058	0.433
Denmark	**0.994**	**0.050**		1.127	**0.090**
Germany	**0.991**	**0.046**		1.355	**0.090**
Canada	**0.962**	**0.000**		2.092	**0.005**
Japan	**0.991**	**0.071**		1.441	0.121
Australia	**0.981**	**0.000**		2.151	**0.009**
Switzerland	**0.993**	0.123		0.955	0.239
Korea	**0.989**	**0.014**		1.645	**0.047**
Norway	**0.997**	0.145		0.511	0.189
Sweden	**0.991**	**0.043**		1.794	**0.044**

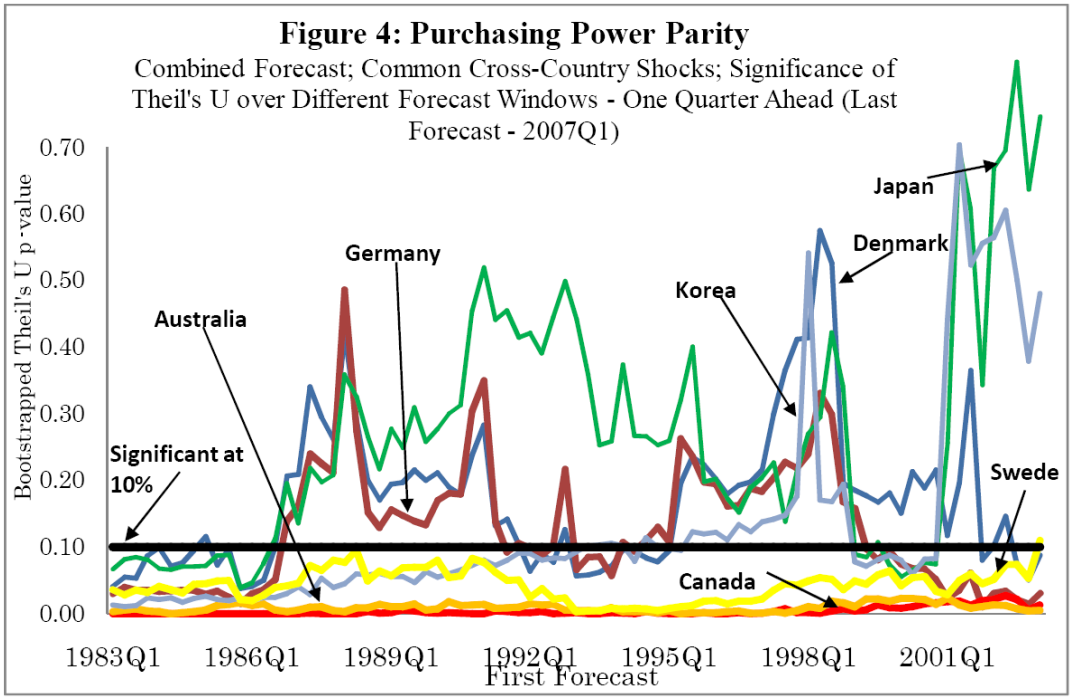
See note of Table 4; Forecasts are calculated as 0.2 times the forecast of the Structural PPP Model which incorporates common cross-country shocks

Exploring the results of Table 5, it is interesting to note that when the forecasts are pooled, the bootstrapped TU becomes statistically significant for the same 7 out of 10 countries for which CW was significant prior to pooling (Table 4).[45] Therefore, by simply pooling the forecasts (a decision we choose to make after observing the significance of the bootstrapped CW and the ENC-NEW), we are able to outperform the driftless random walk in as many as 7 out of 10 cases. As a result, we would suggest that forecasters interested in finding the forecast which produces the smallest MSFE should not ignore the bootstrapped CW and ENC-NEW test statistics. They should explore the potential of improving their forecast by weighting if the bootstrapped TU and DMW are not statistically significant but the bootstrapped CW and ENC-NEW are significant. However, one should note that while exogenously imposed weights tend to perform relatively well, assigning fixed weights is an ad hoc and sub-optimal process. As a result, it does not guarantee that the same weight will continue performing well after a potential future structural break for example.

[45]The bootstrapped DMW test statistic is significant in 6 out of 10 cases.

Robustness of the Pooled Forecast with Respect to Different Forecast Windows - Is Pooling Enough?

Finally, we test whether our combined forecasts which incorporate the information presented by the bootstrapped CW and ENC-NEW test statistics are robust over different periods of time. As before, we test the robustness of the results by graphing the bootstrapped TU p-value for different forecast windows. Only those countries for which we have significant results for a number of different forecast windows are included. The "pooled" PPP specification seems to perform exceptionally well in the early-to mid 1980s and relatively well in the early to mid-1990s.



Nevertheless, the only two countries for which there is robust evidence for forecastability are the commodity exporters Canada and Australia since the bootstrapped TU p-value for these countries is always below 0.1 regardless of the forecast window considered.[46] It is interesting to note that the forecasting success we observe for Australia and Canada is a result of the way we specify the time dummy effect forecast and not so much of the economic fundamental we use. Even a specification with no structural fundamental (only time and country dummies) produces relatively robust results for Australia and Canada. While these results might be of interest to

[46]There is no guarantee that the relationship for Australia and Canada will be preserved in the future. As a result, one should be careful when interpreting the results. Furthermore, we do not observe similar success when we use the same approach to forecast the exchange rates of other commodity producers such as New Zealand and South Africa. Note also that the results are somewhat robust for Sweden as well.

forecasters, they would most likely be of lesser value to policy makers who are interested in the relationship between structural models and fundamentals.[47]

Overview and Summary of PPP Model (with Common Cross-Country Shocks)

To summarize, panel forecast techniques should improve our ability to forecast exchange rates by increasing the sample size and by allowing for cross-country interactions. We argue that, to an extent, forecasters can exploit cross-country interactions even further via specifying the time dummy effect forecast in a way which captures world economic trends. However, while allowing for the incorporation of cross-country information produces slight improvement over simple panel specifications, it fails to produce robust results for the majority of the countries considered. The only exceptions are the commodity producers – Canada and Australia – but we caution the reader that further investigation of these "success" cases is required. Last but not least, while alternative ways to forecast the time dummy or pooling the structural model coefficient across countries may potentially improve our ability to forecast exchange rates for some countries, one should be cautious when interpreting the results.[48] If our ability to forecast exchange rates can be attributed solely to "ad hoc" procedures that take into account *unknown* cross-country shocks and common relationships, we still have not improved significantly our knowledge of the relationship between structural models and exchange rates.

Conclusion

In this paper we attempt to answer the question "Are structural models getting closer to being able to forecast exchange rates at short horizons?" and the answer is "A little." However, over-reliance on asymptotic test statistics in out-of-sample comparisons, misinterpretation of some tests, and failure to sufficiently check robustness to alternative time windows has led many studies to overstate even the relatively thin positive results that have been found. We find that by allowing for common shocks in our panel specification, we are able to generate some improvement, but even that improvement is not entirely robust to the forecast window, and much of the gain appears to come from non-structural rather than structural factors.

[47] For example, while our results suggest that common cross-country shocks seem to forecast the exchange rates of Australia and Canada relatively well, this result does not help policy makers determine the cause of these shocks or determine the relationship between structural variables and the exchange rate. A recent paper by Chen, Rogoff and Rossi (2008) is an example of the difficulty of forecasting the exchange rates of commodity producers solely using fundamentals such as commodity prices even when one takes into account structural breaks.

[48] For example, Rapach and Wohar (2004) provide empirical evidence against pooling the monetary model coefficient across countries.

We explore the application of popular new out of sample test statistics such as the Clark and West (2006, 2007) and Clark and McCracken (2001) out-of-sample test statistics. We argue that they have been widely misinterpreted as minimum mean square forecast error test statistics and that, in addition, popular simple asymptotic versions may suffer from size distortions. In other words, significant Clark-West and Clark-McCracken test statistics do not always imply that the forecast of the structural model outperforms the forecast of the random walk in terms of mean square forecast error. For this question, statistics such as the bootstrapped Theil's U or Diebold-Mariano/West may be more appropriate (especially given the advances in time series bootstrapping); at the very least, researchers should test the robustness of their results with respect to alternative test statistics.

We note that some researchers may be specifically interested in whether one can reject the null hypothesis that the true model is the random walk model in favor a particular structural model. But we would argue that in the vast majority of applications, policy-makers and practitioners treat the random walk model only as a straw man, and simply want to know whether the structural model can deliver a better forecast and what that forecast is.

We do note that, in principle, a positive CW statistic implies that there does exist some linear combination of the driftless random walk and the structural model that outperforms the naive random walk as measured by relative mean square forecast error. Finding a stable linear combination, however, is tricky and potentially opens up a whole new range of problems. Endogenous methods for finding optimal weights tend to fail due to the presence of structural instability. In practice, fixed exogenous weights tend to perform better, although here too stability is a challenge.

In addition to misinterpretation of the new out-of-sample tests for nested models, some of the excess optimism in the literature can be attributed to the failure to check for robustness over different forecast windows. Regardless of whether one uses new or old structural models, single equation or panel specifications, one of the main problems related to the forecastability of the majority of exchange rates remains - lack of robustness over different time periods. Whether the lack of robustness is due to non-linear functional forms, structural breaks or simply heterogenous market sentiments over time[49], the literature on exchange rate forecasting has not been able to develop the tools to produce robust forecasts for the majority of exchange rates. Innovative approaches of overcoming these problems are required in order for the forecasts of

[49]One way of explaining the lack of robustness is with the existence of structural breaks which are identified as one of the main problems related to out-of-sample forecasting (see Clements and Hendry, 2005, Rossi, 2005, Stock and Watson, 1996, 2003). Potential model mis-specification could be an alternative explanation. Empirical evidence suggest that the relationship between fundamentals and exchange rates can be better represented by non-linear rather than linear functional forms (see Taylor and Peel, 2000, Meese and Rose, 1991 and Kilian and Taylor, 2001). However, even when forecasters try to account for non-linear functional forms directly (Meese and Rose,1991, and Killian and Taylor, 2001), or estimate a regime switching model (Marsh, 2000, and Dacco and Satchell, 1999), results remain non-robust.

structural models to outperform the forecasts of the driftless random walk at short-horizons. Until then, we would call the glass ninety-five percent empty rather than five percent full.

References

Alquist, R., and M. Chinn (2006), "Conventional and Unconventional Approaches to Exchange Rate Modeling and Assessment", *NBER* Working Paper No.12481.

Ardic, O. P., O. Ergin and G. B. Senol (2008), "Exchange Rate Forecasting: Evidence from the Emerging Central and Eastern European Economies", *mimeo*, Bogazici University.

Basher, S., and J. Westerlund (2006), "Can Panel Data Really Improve the Predictability of the Monetary Exchange Rate Model?", *MPRA* Working Paper No. 1229.

Berkowitz, J., and L. Kilian (1996), "Recent Developments in Bootstrapping Time Series," *FEDS* Discussion Paper No. 96-45.

Boucher, C. (2006), "Stock Prices – Inflation Puzzle and the Predictability of Stock Market Returns", *Economics Letters* 90(2), 205-212.

Brownstone, D., and B. Valletta (2001), "The Bootstrap and Multiple Imputations: Harnessing Increased Computing Power for Improved Statistical Tests", *The Journal of Economic Perspectives* 15(4),129-141.

Cerra, V., and S. C. Saxena (2008), "The Monetary Model Strikes Back: Evidence from the World", *IMF* Working Paper Series WP/08/73.

Chen,Y., K. Rogoff and B. Rossi (2008), "Can Exchange Rates Forecast Commodity Prices?", *mimeo*, Harvard University.

Cheung, Y.-W., M. Chinn and A. Pascual (2003), "Empirical Exchange Rate Models of the Nineties: Are Anyfit to Survive?", *mimeo*, University of California at Santa Cruz.

Chong, Y., and D. Hendry (1986), " Econometric Evaluation of Linear Macro-Economics Models", *Review of Economics Studies* 53, 671-690.

Clark, T., and K. West (2006), "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis", *Journal of Econometrics* 135, 155-186.

Clark, T., and K. West (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models", *Journal of Econometrics* 138, 291-311.

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

Clark, T., and M. McCracken (2005), "Evaluating Direct Multi-Step Forecasts", *Econometric Reviews* 24(4), 369–404.

Clements, M., and D. Hendry (1993), " On the Limitations of Comparing Mean Square Forecast Errors", *Journal of Forecasting* 12, 617-637.

Clements, M., and D. Hendry (1996), "Intercept Corrections and Structural Change", *Journal of Applied Econometrics* 11, 475-494.

Clements, M., and D. Hendry (1998), "Forecasting Economic Time Series", *Cambridge University Press*.

Clements, M., and D. Hendry (1999), "Forecasting Non-Stationary Economic Time Series", *Cambridge University Press*.

Clements, M., and D. Hendry (2004), "Pooling of Forecast", *Econometrics Journal* 7(1), 1-31.

Clements, M., and D. Hendry (2005), "Forecasting with Breaks", in Elliott, G., C. Granger and A. Timmermann (ed.), *Handbook of Economic Forecasting*, Elsevier, 1(1),1, 605-657 (2006).

Dacco, R., and S. Satchell (1999), “Why Do Regime-Switching Models Forecast So Badly?”, *Journal of Forecasting* 18, 1-16.

Diebold, F., and R. Mariano (1995), “Comparing Predictive Accuracy”, *Journal of Business and Economic Statistics* 13, 253-265.

Efron, B. (1979), “Bootstrap Methods: Another Look at the Jackknife,” *Annals of Statistics* 9, 1218-1228.

Engel, C. (1996), "The Forward Discount Anomaly and the Risk Premium: A Survey of Recent Evidence", *Journal of Empirical Finance* 3, 123–191.

Engel, C., and J. Hamilton (1990), "Long Swings in the Exchange Rate: Are They in the Data and Do Markets Know It?", *American Economic Review* 80 (4), 689-713.

Engel, C. and K. West (2005), "Exchange Rates and Fundamentals," *Journal of Political Economy* 113 (3), 485-517.

Engel, C., Mark, N. and K. West (2007), "Exchange Rate Models Are Not as Bad as You Think," *NBER Macroeconomics Annual*.

Ericsson, N. (1992), "Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration," *Journal of Policy Modeling* 14 (4), 465-495.

Franses, P., and R. Legerstee (2007), "Does Experts Adjustment to Model-Based Forecasts Contribute to Forecast Quality?" *Econometric Instititute Report* 37.

Friedman, B., and K. Kuttner (1992), "Money, Income and Prices After the 1980s," *NBER* Working Papers 2852.

Giacomini, R., and B. Rossi (2006), "How Stable is the Forecasting Performance of the Yield Curve for Output Growth?" *Oxford Bulletin of Economics and Statistics* 68, 783-795.

Giacomini, R., and B. Rossi (2008), "Forecast Comparisons in Unstable Environments", *mimeo,* Duke University.

Gourinchas, P.-O., and H. Rey (2007). "International Financial Adjustment," *Journal of Political Economy* 115, 4.

Groen, J. (2005), "Exchange Rate Predictability and Monetary Fundamentals in a Small Multi-Country Panel", *Journal of Money, Credit and Banking* 37, 495-516.

Groen, J. (2007), "Fundamentals Based Exchange Rate Prediction Revisited", *mimeo*, Federal Reserve Bank of New York.

Harvey,D., and P. Newbold (1998), " Tests for Forecast Encompassing." *Journal of Business and Economic Statistics* 16, 2.

Holden, K., and D. A. Peel (1989), "Unbiasedness, Efficiency and the Combination of Economic Forecasts." *Journal of Forecasting* 8, 175-188.

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.

Li, H. and Maddala, G. (1997), "Bootstrapping Cointegrating Regressions", *Journal of Econometrics*, 80 (2), 297-318.

Maddala, G. and I.-M. Kim (1998), *Unit Roots, Cointegration and Structural Change*, Cambridge University Press.

MacKinnon,J. (2002), "Bootstrap Inference in Econometrics", *Canadian Journal of Economics*, 35 (4), 615-645.

Marcellino, M. (2000), "Forecast Bias and MSFE encompassing", *Oxford Bulletin of Economics and Statistics*, 62, 533-542.

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

Mark, N., and D. Sul (2001), "Nominal Exchange Rates and Monetary Fundamentals: Evidence from a Small Post-Bretton Woods Sample", *Journal of International Economics*, 53, 29-52.

Marsh, I. (2000), "High-Frequency Markov Switching Models in the Foreign Exchange Market", *Journal of Forecasting* 19, 123-134.

McCracken, M. (1999), "Asymptotics for out of sample tests of causality", *mimeo*, University of Missouri.

McCracken, M., and S. Sapp (2005), "Evaluating the Predictability of Exchange Rates Using Long Horizon Regressions: Mind Your p's and q's!", *Journal of Money, Credit and Banking* 37, 473-494.

Meese, R., 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: Sampling Error or Misspecification," in Jacob A. Frenkel, (ed.), *Exchange Rates and International Macroeconomics*, NBER.

Meese, R., and A. Rose (1991), "An Empirical Assessment of Non-Linearities in Models of Exchange Rate Determination", *The Review of Economic Studies* 58 (3), 603-619.

Molodtsova, T., and D. Papell (2008), "Out-of-Sample Exchange Rate Predictability with Taylor Rule Fundamentals", *mimeo*, University of Houston.

Molodtsova,T., A. Nikolsko-Rzhevskyy and D. Papell (2007), "Taylor Rules with Real-Time Data: A Tale of Two Countries and One Exchange Rate," *mimeo*, University of Houston.

Molodtsova,T., A. Nikolsko-Rzhevskyy and D. Papell (2008), "Taylor Rules and the Euro," *mimeo*, University of Houston.

Politis, D., and H. White (2004), "Automatic Block-Length Selection for the Dependent Bootstrap," *Econometric Reviews* 23(1), 53 - 70.

Rapach,J., M. Wohar and D. Rangvid (2005), "Macro variables and international stock return predictability", *International Journal of Forecasting* 21 (1), 137-166.

Rapach, D., J. Strauss and M. Wohar (2007), "Forecasting Stock Return Volatility in the Presence of Structural Breaks", in David E. Rapach and Mark E. Wohar (eds.), *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, Amsterdam: Elsevier, forthcoming.

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.

Rapach, D., and M. Wohar (2004), "Testing the Monetary Model of Exchange Rate Determination: A Closer Look at Panels", *Journal of International Money and Finance* 23, 867-895.

Reinhart, C., and K. Rogoff (2002), "The Modern History of Exchange Rate Arrangements: A Reinterpretation", *NBER* Working Paper.

Rossi, B. (2006), "Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability", *Macroeconomic Dynamics* 10(1), 20-38.

Sellin, P. (2006), "Using a Real Exchange Rate Model to Make Real and Nominal Forecasts", *mimeo*, Sveriges Riksbank.

Stock, J., and M. Watson (1996), " Evidence of Structural Instability in Macroeconomic Time Series Relations", *Journal of Business and Economic Statistics* 14, 11-30.

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

Taylor, M., and D. Peel (2000), "Nonlinear adjustment, long-run equilibrium and exchange rate fundamentals", *Journal of International Money and Finance* 19(1),33-53.

Zagaglia, P. (2006), "The Predictive Power of the Yield Spread under the Veil of Time", *Stockholm University*, Research Papers in Economics No 4.

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

APPENDIX:

A. Minimum MSFE Out-of-Sample Test Statistics

The Theil's U Test (TU)

The TU test statistic is a minimum MSFE test defined as the square root of the MSFE of the structural model over the square root of the MSFE of the random walk model. *Therefore, a TU $<$ 1 implies that the structural model outperforms the random walk model.* TU is often preferred for its simplicity and intuitive interpretation and its statistical significance is tested via a bootstrap. The test we use is a one-sided test.

The Diebold – Mariano/West Test (DMW)

The DMW test statistic can be considered an alternative to the TU test. It measures the statistical significance of the difference between the MSFE of the random walk model and that of the structural model. *A significant and positive DMW test implies that the structural model outperforms the random walk.* On the basis of both theoretical and simulation evidence, West (1996), McCracken (1999), Clark and McCracken (2001, 2005) and Clark and West (2006) show that, when comparing nested models, the asymptotic DMW test statistic is undersized, which means that it may not detect statistical significance (i.e., that the structural model outperforms the random walk model) even when it exists. While Clark and West (2006) attribute the poor size of the asymptotic DMW test statistic to small-sample bias, McCracken (1999) and Clark and McCracken (2001, 2005) claim that the asymptotic DMW is undersized because the limiting distribution of the DMW under the null hypothesis is not standard normal when nested models are compared. To correct for this problem, a number of studies opt for the bootstrapped DMW test statistic which does not assume any distributional form. This is the approach we take in this paper. Again, we calculate the DMW test as a one sided test.

The Clark – West Test (CW)

To compensate for the fact that the asymptotic DMW test statistic is undersized under the null hypothesis when comparing nested models and to avoid the use of a bootstrap, Clark and West (2006, 2007) propose a new asymptotic test for nested models, the CW, that builds on the asymptotic DMW test. The CW test statistic takes into account the fact that the two models compared are nested by assuming that, under the null hypothesis, the exchange rate follows a random walk.

When the forecast is calculated using rolling regressions, the limiting distribution of the CW under the null hypothesis is standard normal. However, when the estimation is performed recursively, the asymptotic distribution is approximated using Brownian motion. Based on simulation evidence, Clark and West (2007) suggest that, for recursive specifications, one can use a one-sided test and should reject the null hypothesis of equal forecasting power when the CW $\geqq +1.282$ at 10 percent and CW $\geqq +1.645$ at 5 percent (which are the critical values one would use assuming a normal distribution). Finally, we also calculate the bootstrapped CW test statistic to test whether the asymptotic CW test is properly sized.

The Clark - McCracken Test (ENC-NEW)

Another relatively new out-of-sample test statistic for nested models, the ENC-NEW, introduced by Clark and McCracken (2001, 2005), also implicitly assumes that, under the null, the exchange rate follows a random walk.[50] The ENC-NEW and the CW differ only by a scaling factor. In other words, the two test statistics can differ slightly because of different power or size but they test the same null hypothesis. The shortcoming of the ENC-NEW is that its asymptotic distribution is a function of both the in-sample and out-of-sample portion of the data which makes evaluation of statistical significance quite cumbersome. Therefore, bootstrapping the ENC-NEW is an attractive alternative and this is the approach we take in this paper.

B. Proofs: The New Out-of-Sample Tests for Nested Models

In this section of the Appendix we provide a theoretical argument why the CW and ENC-NEW cannot be considered minimum MSFE tests in cases of severe "scale" forecast bias. We also argue that one should interpret these out-of-sample tests for nested models as prior tests of whether one can pool the random walk and the structural model forecast to produce a forecast with MSFE significantly smaller than the MSFE of the random walk.

[50] Similarly to the CW, the ENC-NEW has been one of the most widely used out-of-sample test statistics in the exchange rate forecasting literature. Some of the studies that test out-of-sample forecastability using the ENC-NEW are Franses and Legerstee (2007), Rossi (2006), Zagaglia (2006), Boucher (2006), Giacomini and Rossi (2006), Rangvid, Rapach, Wohar (2005).

From the Diebold – Mariano to the Clark – West Test

Before proceeding to the proofs, we present a derivation of the Clark and West test statistic, as presented in Clark and West (2006). In order to simplify the notation, assume that the forecast is one period ahead and that the forecast variable is the change in the exchange rate. Assume that $y_t = s_t - s_{t-1}$, where s_t is the natural log of the exchange rate for period t. Also let X_t be a matrix of explanatory variables. We are interested in comparing the forecasting power of the following theoretical models:

$$\begin{aligned} \text{Driftless Random Walk Model:} \quad & y_t = e_{1,t}, \text{ and} \\ \text{Structural Model:} \quad & y_t = X_{t-1}b + e_{2,t}, \end{aligned}$$

where $e_{1,t}$ and $e_{2,t}$ are the unobservable innovation terms.

The CW test assumes that, under the null hypothesis, the exchange rate is a random walk, and therefore, the population parameter $b = 0$, and the forecast innovation terms are equal, that is, $e_{1,t+1} = e_{2,t+1}$. The models can be estimated by OLS using either recursive or rolling regressions. The estimated forecasts for the random walk and the structural model are $\hat{y}_{1,t+1} = 0$ and $\hat{y}_{2,t+1} = X_t\hat{b}_t$ respectively. Denoting P as the number of forecasts, T as the sample length, and R as the sample reserved to calculate the first forecast, we can rewrite the sample difference between the MSFE of the two models (which is the main component of the DMW test statistic) as:

$$P^{-1}\sum_{t=R+1}^{t=T}\hat{e}_{1,t+1}^2 - P^{-1}\sum_{t=R+1}^{t=T}\hat{e}_{2,t+1}^2 = 2P^{-1}\sum_{t=R+1}^{t=T}(y_{t+1}X_t\hat{b}_t) - P^{-1}\sum_{t=R+1}^{t=T}(X_t\hat{b}_t)^2.$$

Clark and West (2006, 2007) argue that under the null hypothesis $e_{1,t+1} = e_{2,t+1} = y_{t+1}$; and since Clark and West (2006, 2007) assume that the independent variables are not correlated with the theoretical disturbance terms, it follows that $E(y_{t+1}X_t\hat{b}_t) = 0$.[51] Therefore, they argue that we should expect $\sum_{t=R+1}^{t=T}(y_{t+1}X_t\hat{b}_t) \approx 0$ for both the rolling and recursive specifications. However, due to small-sample bias, $-P^{-1}\sum_{t=R+1}^{t=T}(X_t\hat{b}_t)^2 < 0$. As a result, the sample difference of the MSFEs of the random walk and the structural model is negatively biased in favor of the random walk.

The fact that the DMW test is negatively biased under the null hypothesis implies that it favors the random walk. Therefore, Clark and West (2006) propose an "adjusted" DMW statistic, or the so-called CW statistic, which tests whether

[51] $e_{1,t+1} = e_{2,t+1}$ implies $E(y_{t+1}X_t\hat{b}_t) = E(e_{1,t+1}X_t\hat{b}_t) = E(e_{2,t+1}X_t\hat{b}_t)$. By assumption, $E(e_{2,t+1}X_t) = 0$. Then if one assumed that underlying variables are independent, $E(e_{2,t+1}X_t\hat{b}_t) = E(e_{2,t+1}X_t)E(\hat{b}_t) = 0$.

$$\hat{d} = 2P^{-1} \sum_{t=R+1}^{t=T} (y_{t+1} X_t \hat{b}_t) \tag{6}$$

is significantly greater than zero. If it is, the structural model outperforms the random walk. More formally, we can define the CW as

$$CW = \frac{P^{0.5}\hat{d}}{\sqrt{\Omega^{\hat{d}}}}$$

where $\Omega^{\hat{d}}$ is the variance of $\hat{d}$. In comparison, one can write the ENC-NEW test statistics as

$$ENC - NEW = \frac{P \sum_{t=R+1}^{t=T} (y_{t+1} X_t \hat{b}_t)}{\sum_{t=R+1}^{t=T} (y_{t+1} - X_t \hat{b}_t)^2}$$

which makes it clear that the CW and ENC-NEW differ only by a scaling factor and we would expect that they behave similarly.

New Out-of-Sample Test Statistics (CW and ENC-NEW) : Not Minimum MSFE Tests

While the CW (and to a lesser degree the ENC-NEW) are often used interchangeably with the older minimum MSFE tests, we argue that their use as minimum MSFE tests is based on a misinterpretation, and that they should not be used as a substitute for the TU and the DMW tests. We present a theoretical proof that the CW is not a minimum MSFE test.[52] The proof can be easily generalized for the ENC-NEW out-of-sample test statistic.

Our proof is based on the rolling window specification ($R/P \to 0$ and R is fixed) which generalizes well the main point made by Clark and West (2006, 2007), namely, the presence of small-sample bias. The proof for the recursive case is similar.[53] While the proof assumes that the benchmark model is the driftless random walk, it can be generalized to any nested-model specification.

[52] Clark and West themselves suggest that researchers should interpret the CW as a minimum MSFE test statistic (2007, pp. 297). Clark and McCracken (2001, 2005) do not make such a claim regarding the ENC-NEW.

[53] The recursive specification can be analyzed in a framework either with or without small-sample bias. If we assume that small sample bias is present, even when $\hat{b}_t$ is estimated using recursive regressions, as Clark and West (2006) argue, we would still expect under certain assumptions that $2P^{-1} \sum_{t=R+1}^{t=T} (y_{t+1} X_t \hat{b}_t) \approx 2E(y_{t+1} X_t \hat{b}_t)$. As a result, in the presence of small-sample bias, the proof we present generalizes to the recursive case.

In the case when $R \to \infty, P \to \infty$ (i.e., no small-sample bias is present), the proof we present still holds. However, this case is irrelevant, given that according to Clark and West's (2006, 2007) null hypothesis, if small-sample bias was not an issue, the adjustment of the DMW the authors propose would not be justified (under the null, the negative bias would disappear since $b = 0$ and, as a result, $E(X_t b)^2 = 0$).

Assume that all the variables are defined as above. In the rolling regression specification, the null hypothesis incorporates the presence of small-sample bias and can be defined as $Ee^2_{1,t+1} = Ee^2_{2,t+1}(R)$. The alternative can be defined as $Ee^2_{1,t+1} > Ee^2_{2,t+1}(R)$. The respective MSFEs are

$$Ee^2_{1,t+1} = E(y_{t+1})^2, \text{ and} \tag{7}$$

$$Ee^2_{2,t+1}(R) = E(y_{t+1} - X_t\hat{b}_t)^2 = E(y_{t+1})^2 - 2Ey_{t+1}X_t\hat{b}_t + E(X_t\hat{b}_t)^2, \tag{8}$$

where $x(R)$ implies that the variable x is a function of a rolling window of fixed size, R. In the rolling regression case (the extreme version of small-sample bias with respect to the structural model parameter), R is fixed, and $\hat{b}_t$ never converges to b, regardless of the sample size, T. The larger R is, the smaller the small-sample bias is. Also, when R is fixed, then under mild assumptions, we would expect $y_{t+1}X_t\hat{b}_t$ to be a well-behaved *iid* random variable. Then, it follows that

$$p\lim(P^{-1}2\sum_{t=R+1}^{t=T}(y_{t+1}X_t\hat{b}_t)) = 2E(y_{t+1}X_t\hat{b}_t),$$

where the probability limit is defined with respect to $P \to \infty$. Given the rolling regression set-up, we proceed to prove that CW is not a minimum MSFE test statistic. In other words, a statistically significant CW test does not imply a statistically significant minimum MSFE test.

Proposition 1: $2E(y_{t+1}X_t\hat{b}_t) > 0 \nRightarrow Ee^2_{1,t+1} - Ee^2_{2,t+1}(R) > 0.$

Proof of Proposition 1: From equations (7) and (8), if $2E(y_{t+1}X_t\hat{b}_t) \leq E(X_t\hat{b}_t)^2$ then $Ee^2_{1,t+1} - Ee^2_{2,t+1}(R) \leq 0$. However $2E(y_{t+1}X_t\hat{b}_t) \leq E(X_t\hat{b}_t)^2$ can hold even if $2E(y_{t+1}X_t\hat{b}_t) > 0$. As a result, $2E(y_{t+1}X_t\hat{b}_t) > 0$ does not imply $Ee^2_{1,t+1} - Ee^2_{2,t+1}(R) > 0$. □

The question emerges how often we would expect the CW(ENC-NEW) and DMW(TU) to produce different results due to the fact that the two test statistics test a different null hypotheses. In other words, how often we would observe $0 < 2E(y_{t+1}X_t\hat{b}_t) \leq E(X_t\hat{b}_t)^2$ in practice. The condition $0 < 2E(y_{t+1}X_t\hat{b}_t) \leq E(X_t\hat{b}_t)^2$ implies that if we regress the observed exchange rate change on the structural model forecast and no constant, the estimated coefficient should be less than or equal to $\frac{1}{2}$ and greater than 0. This is equivalent to having a significantly biased forecast (if the forecast is unbiased the estimated coefficient should be 1).[54]

[54]Note that the analysis refers only to "scale" bias since no constant is included in the forecast bias regression (for details see Marcellino, 2000, and Holden and Peel, 1989).

Forecast bias is a significant problem in the literature on exchange rate forecasting. Marcellino (1998) emphasizes the importance of taking into account forecast bias when applying encompassing test statistics. Clements and Hendry (1996, 2005) investigate the theoretical relationship between structural breaks and forecast bias and find that structural breaks, which are fairly common in forecasting, can lead to forecast bias.

What Do the New Out-of-Sample Test Statistics for Nested Models Test?

Here we prove that, *in theory*, a significant CW test implies that one can pool the random-walk and the structural-model forecasts and obtain a combined forecast whose MSFE is smaller than that of the random walk. Again, this proof also applies to the ENC-NEW test statistic.

Similarly to the proof of Proposition 1, we prove the statement above in the context of the rolling specification, which implies that there is small-sample bias under the null hypothesis. However, a similar proof can be presented with respect to the recursive specification. The proof here is generalized to any nested model specification where Model 1 is nested in Model 2. Let $y_{c,t+1} = \lambda y_{2,t+1}(R) + (1-\lambda)y_{1,t+1}(R), 0 \leq \lambda \leq 1$ be the combined forecast where λ is the weight on the structural model forecast. Subscripts represent the respective model (1 or 2) and c stands for "combined". As before, the variable that we forecast is $y_{t+1.}$ One can rewrite the CW test statistic as testing whether

$$\hat{d} = 2P^{-1} \sum_{t=R+1}^{t=T} \hat{e}_{1,t+1}(\hat{e}_{1,t+1} - \hat{e}_{2,t+1})$$

is significantly greater than zero. Within the more general framework of any nested model specification, we prove that a significant CW implies that there exists an optimal combination between the two forecasts which will produce a combined forecast that outperforms the simpler model (Model 1) in terms of MSFE.

$$\begin{aligned} \textit{Proposition 2}: \quad & 2Ee_{1,t+1}(R)(e_{1,t+1}(R) - e_{2,t+1}(R)) > 0 \\ \Rightarrow \quad & \exists \ \lambda \ s.\ t.\ Ee^2_{1,t+1}(R) - Ee^2_{c,t+1}(R) > 0 \end{aligned}$$

Proof of Proposition 2:[55] The proof we present is similar to the proof provided by Harvey and Newbold (1998). In a rolling regression framework and under mild assumptions

$$p\lim(2P^{-1}\sum_{t=R+1}^{T}\hat{e}_{1,t+1}(\hat{e}_{1,t+1}-\hat{e}_{2,t+1})) \quad = \\ 2Ee_{1,t+1}(R)(e_{1,t+1}(R)-e_{2,t+1}(R)), P \quad \rightarrow \quad \infty$$

We can minimize the MSFE of the combined forecast, by regressing y_{t+1}, the observed series, on $y_{1,t+1}(R)$ and $y_{2,t+1}(R)$ using OLS and constraining the coefficients to sum to one.

$$y_{t+1} = \lambda y_{2,t+1}(R) + (1-\lambda)y_{1,t+1}(R) + e_{c,t+1}(R), 0 \leq \lambda \leq 1 \tag{9}$$

If $\lambda > 0$, then combining the forecasts will produce a forecast s.t. $Ee^2_{1,t+1}(R) > Ee^2_{c,t+1}(R)$. Equation (9) can be rewritten as

$$e_{1,t+1}(R) = \lambda(e_{1,t+1}(R) - e_{2,t+1}(R)) + e_{c,t+1}(R), \tag{10}$$

If we estimate equation (10) without a constant, then

$$\lambda = \frac{Ee_{1,t+1}(R)(e_{1,t+1}(R)-e_{2,t+1}(R))}{E(e_{1,t+1}(R)-e_{2,t+1}(R))^2}$$

Testing whether $2Ee_{1,t+1}(R)(e_{1,t+1}(R)-e_{2,t+1}(R)) = 0$ is testing the same hypothesis as testing whether $\lambda = 0$ using equation (9) or (10). Therefore, $2Ee_{1,t+1}(R)(e_{1,t+1}(R)-e_{2,t+1}(R)) > 0 \Rightarrow \exists\ \lambda\ s.\ t.\ Ee^2_{1,t+1}(R) - Ee^2_{c,t+1}(R) > 0$ □

As a result, while the CW and the ENC-NEW out-of-sample test statistics cannot be considered minimum MSFE test statistics, they still provide meaningful information. They can be used as a prior test of whether a combined forecast exists that outperforms the driftless random walk forecast in terms of MSFE.

[55] One can also think of CW and ENC-NEW in the framework of encompassing test statistics. If one fails to reject the null that $\hat{d}$ is equal to zero, then the random walk encompasses the structural model. If one rejects the null that $\hat{d}$ is equal to zero, then the CW test statistic is statistically significant and the random walk fails to encompass the structural model. Note that a significant CW test statistic *does not* necessarily imply that the structural model encompasses the random walk. The distinction is important. If the structural model encompasses the random walk, which would occur if we fail to reject the null that $Ee_{2,t+1}(e_{2,t+1}-e_{1,t+1})$ equals 0, then the structural model will have a smaller MSFE than the random walk. As a result, encompassing will entail MSFE dominance (for proof see Ericsson, 1992, and Marcellino,2000).

C. The Bootstrap

Different Bootstrapping Procedures

First, we briefly discuss alternative approaches to bootstrapping. The standard bootstrap (also known as "case" bootstrap with replacement), introduced by Efron (1979), assumes that the re-sampled data are independent and identically distributed (iid). If the data are serially correlated or if heteroskedasticity is present (which is usually the case in time series data), a simple "case" bootstrap leads to inconsistent results. A better way to bootstrap time series data would be via a block bootstrap which implies that the data is re-sampled in blocks of a certain length rather than observation by observation, thereby preserving the properties of the data generating process (DGP). However, finding the optimal block size to preserve the DGP is not that straight forward. Indeed, Berkowitz and Kilian (1996) suggest that block bootstrapping is not the optimal way to bootstrap time series data given the state of development of the block bootstrap literature.

As a better alternative, Berkowitz and Kilian (1996) suggest a residual bootstrap procedure which is the bootstrap specification implemented by Mark (1995) and subsequently improved by Kilian (1999) and Mark and Sul (2001). The idea of residual bootstrapping is that in specifications such as Mark (1995) where the independent variable is defined as the deviation of the exchange rate from the fundamental, the cointegration (or the lack of cointegration) will be preserved when one uses the residual bootstrap. One way to implement it is to estimate an error correction specification, then re-sample the estimated residuals and recursively simulate the independent variable. (This type of bootstrap is commonly referred to as "semi-parametric" bootstrap. If one draws the residuals from a normal distribution, the bootstrap will be called "parametric".) While not always easy to implement, if properly specified, the bootstrap automatically corrects for small-sample bias and can be also used for forecast horizons greater than one as discussed in Kilian (1999).

Bootstrap Used in the Paper

The bootstrap procedure used to calculate the p-values of DMW, TU, CW and ENC-NEW for all specifications is similar to the bootstrap of Mark and Sul (2001) and Basher and Westerlund (2006). The main difference between our bootstrap and Mark and Sul's (2001) bootstrap is that we use country specific OLS - regressions rather than seemingly unrelated regressions (SURs). We also perform a "semi-parametric" rather than "parametric" bootstrap. The data generating process (DGP) is a country – specific error correction process. The assumption of no unit root (or cointegration between the fundamental and the exchange rate) is imposed.

For each country, we estimate the following equations using an OLS regression (the "i" subscript is dropped for simplicity) :

$$\triangle s_t = \varepsilon_t^s$$

$$\Delta z_t = \mu + t + \gamma z_{t-1} + \sum_{k=1}^{d} \delta_k \Delta s_{t-k} + \sum_{k=1}^{l} \zeta_k \Delta z_{t-k} + \varepsilon_t^z$$

where z_t is the deviation of the exchange rate from the fundamental (or simply the fundamental) defined in (2), (3), (4) and (5). s_t is the nominal exchange rate. We also define $\Delta s_t = s_t - s_{t-1,}$ $\Delta z_t = z_t - z_{t-1}$, μ is a constant and t is a trend.

Lags of Δs_t and Δz_t are included in the error correction equation to account for potential autocorrelation. The bootstrap procedure uses the Akaike's information criterion (AIC) method in order to choose between the appropriate number of lags of Δs_t and Δz_t (d and l can differ). The DGP can also differ across countries depending on whether AIC picks a specification with no constant, with constant or with a constant and a trend. The restriction that the sum of the coefficients of the lags of Δz_t equals one is imposed in order to avoid exploding simulated series. Then we re-sample the estimated matrix of residuals, $(\varepsilon^s, \varepsilon^z)$, either case by case (or more precisely row by row) or in blocks of 4 for quarterly data and 12 for monthly data. The results are relatively robust to the alternative methods of re-sampling of the residuals. Therefore, the results presented in the paper are based on case by case re-sampling of the residuals rather than block re-sampling.

Once the residuals are re-sampled, the exchange rate and the independent variable(s) are simulated recursively. The first 100 simulated observations are discarded in order to attenuate potential bias related to choosing the starting values of the recursion - the sample averages. With the new generated sample, the forecasting model is re-estimated and the test statistics are calculated. The p-values of the DMW, CW and ENC-NEW test statistics are measured as the portion of the distribution above the test statistics estimated using the observed data (since all these tests are one-sided tests), while the p-value of the TU statistic is the proportion of the bootstrapped TU distribution below the estimated TU value using the observed data. All the bootstrapped p-values are calculated on the basis of 1000 simulated distributions.