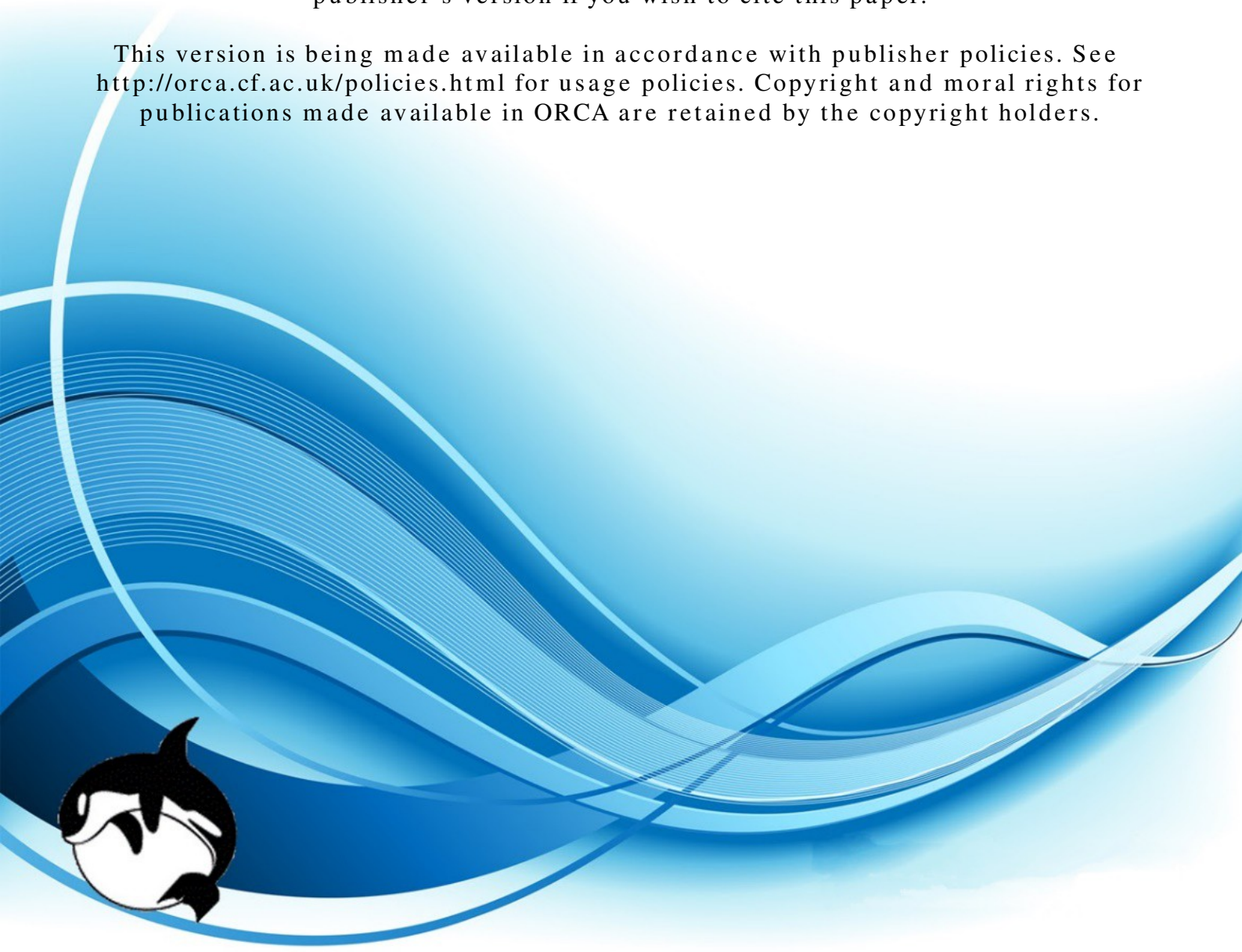

How good are out of sample forecasting Tests on DSGE models?*

Patrick Minford(Cardiff University and CEPR)†
Yongdeng Xu(Cardiff University)‡
Peng Zhou(Cardiff Metropolitan University)§

Abstract

Out-of-sample forecasting tests of DSGE models against time-series benchmarks such as an unrestricted VAR are increasingly used to check a) the specification and b) the forecasting capacity of these models. We carry out a Monte Carlo experiment on a widely-used DSGE model to investigate the power of these tests. We find that in specification testing they have weak power relative to an in-sample indirect inference test; this implies that a DSGE model may be badly mis-specified and still improve forecasts from an unrestricted VAR. In testing forecasting capacity they also have quite weak power, particularly on the lefthand tail. By contrast a model that passes an indirect inference test of specification will almost definitely also improve on VAR forecasts.

JEL categories: E10, E17
Key Words: Out of sample forecasts, DSGE, VAR, specification tests, indirect inference, forecast performance

* We are grateful to participants in the 2014 Konstanz Seminar on Monetary Theory and Policy for discussions of an early contribution to these issues; also to an anonymous referee for most useful comments on an earlier version of this paper.
† Cardiff Business School, Cardiff University, Colum Drive, Cardiff, CF10 3EU, UK.
‡ Corresponding author: Cardiff Business School, Cardiff University, Colum Drive, Cardiff, CF10 3EU, UK.. Email: xuy16@cardiff.ac.uk.
§ Cardiff School of Management, Cardiff Metropolitan University, Cardiff, CF5 2YB, UK.

1. Introduction

In recent years macro-economists have turned to out-of-sample forecasting (OSF) tests of Dynamic Stochastic General Equilibrium (DSGE) models as a way of determining their value to policymakers both for deciding policy and for improving forecasts. Thus for example Smets and Wouters (2007) showed that their model of the US economy could beat a Bayesian Vector Auto Regression (VAR) or BVAR, their point being that while they had estimated the model by Bayesian methods with strong priors there was a need to show also that the model could independently pass a (classical specification) test of overall fit, otherwise the priors could have dominated the model's posterior probability. Further papers have documented models' OSF capacity, including Gürkaynak et al (2013); see Wickens (2014) for a survey of recent attempts by central banks to evaluate their own DSGE models' OSF capacity[1]. But how good are these OSF tests? This question is what this paper sets out to answer.

The value of DSGE models' OSF capacity to policymakers comes as we said from two main motivations.

The first is to use DSGE models to improve economic forecasting. One can think of an unrestricted VAR as a method that uses data to forecast without imposing any theory. Then if one knows the true theory one can improve the efficiency of these forecasts by imposing this theory on the VAR, to obtain the restricted VAR. This will improve the forecasts, reducing the Root Mean Square Error (RMSE) of forecasts at all horizons. However imposing a false parameter structure on the VAR may produce worse forecasts; the further from the truth the parameters are the worse the forecasts. There will be some 'cross-over point' along this falseness spectrum at which the forecasts deteriorate compared with the unrestricted VAR.

The second reason is the desire to have a well-specified model that can be used reliably in policy evaluation; clearly in assessing the effects of a new policy the better-specified the model, the closer it will get to predicting the true effects. The assessment of the DSGE model's forecasting capacity is being used by policymakers with this desire, as a means of evaluating the extent of the model's mis-specification.

[1] Other papers that have computed OSF performance of DSGE models relative to time-series models include: Adolfson, Linde and Villani (2007), Edge and Gürkaynak (2010), Edge, Kiley and Laforte (2010), Giacomini and Rossi (2010), and Del Negro and Schorfheide (2012).

Notice that the two motivations are linked by the requirement of a well-specified model. Thus for the DSGE model to give better forecasts than the unrestricted VAR it needs to be not too far from the true model- i.e. the right side of the cross-over point. It is harder for us to judge how close the model needs to be to the truth for a policy evaluation: this will depend on how robust the policy is to errors in its estimated effects and this will vary according to the policy in question. But we can conclude that both reasons require us to be confident about the model's specification.

Thus evaluations of the DSGE model's forecasting capacity, to be useful, should provide us with a test of the model's specification; and this indeed is how these evaluations are presented to us. Typically the model's forecasting RMSE is compared with that of an unrestricted VAR, e.g. the ratio of the model's RMSE to that of the VAR; there is a distribution for this ratio for the sample size involved and we can see how often the particular model's forecasts give a ratio in say the 5% tail, indicating model rejection. The asymptotic distribution for this ratio (of two t-distributions) cannot be derived analytically; but we establish below by numerical methods that it is a t-distribution.

The questions we ask in this paper are:

- what is the small sample distribution for this ratio for a model 1) if it is true and 2) if it is marginally able to improve other forecasts?
- how much power do these OSF evaluations have, viewed as a test of a DSGE model's specification? In other words can we distinguish clearly between the forecasting performance of a badly mis-specified model and the true model.
- can we say anything about the relationship between a DSGE model's degree of mis-specification and its forecasting capacity? There is a large literature on forecast success of different sorts of models- Clements and Hendry (2005); Christoffel, Coenen and Warne (2011). We would like to see how success is related to specification error.

We investigate these questions using Monte Carlo experiments for a model of the DSGE type being evaluated here; we do so using sample sizes for the out-of-sample forecasts that are of the same order as those used in these tests and so rely not on the asymptotic but on the small sample distributions of the models. In section 2 that follows we explain the OSF tests of a DSGE model. In section 3 we set out the Monte Carlo experiments and show the power of OSF tests of a DSGE model's specification.

In section 4 we establish some links between a DSGE model's specification error and its capacity to improve forecasts. Section 5 concludes.

2. DSGE models out-of-sample forecasting tests

2.1 DSGE model OSFs

A DSGE model (e.g. that of Smets and Wouters, 2007, henceforth SW)) has a general form:

$$\begin{aligned} A_0 E_t(y_{t+1}) &= A_1 y_t + B_0 z_t \\ z_{t+1} &= R z_t + \varepsilon_{t+1} \end{aligned} \tag{1}$$

where y_{t+1} are endogenous variables, z_t are exogenous variables, typically errors, which may be represented by an autoregressive process in which ε_{t+1} are shocks (i.e. $NID(0,\Sigma)$). The solution to a DSGE model can be represented by a restricted VAR:

$$x_{t+1} = Ax_t + B\varepsilon_{t+1} \tag{2}$$

where $x_{t+1} = (y_{t+1}, z_{t+1})'$. The coefficient matrices A and B are full rank but restricted.

A and B can be derived analytically (see Wickens, 2014). Alternatively, if we input the parameter set $\Omega = \{A_0, A_1, B_0, R\}$ into the programme Dynare (Juilliard, 2001), then A and B in (2) can be derived by it. OSFs are then derived straightforwardly from (2). Suppose the initial forecast origin is m, then the OSFs are:

$$\begin{aligned} \hat{x}_{m+1} &= Ax_m \\ \hat{x}_{m+2} &= A\hat{x}_{m+1} = A^2 x_m \\ &\cdots \\ \hat{x}_{m+l} &= A\hat{x}_{m+l-1} = A^l x_m \end{aligned} \tag{3}$$

where $l = 1, 2, \ldots h$. $\hat{x}_{m+l}$ denotes the l -step ahead forecast. We also create False models whose parameters are altered from those of the True one in a manner we explain below.

2.2 VAR model OSFs

Consider the first order VAR

$$y_{t+1} = Py_t + \varepsilon_{t+1} \tag{4}$$

where ε_t is assumed to be $NID(0,\Sigma)$. Suppose the initial forecast origin is m, the OSFs are:

$$\begin{aligned} \hat{y}_{m+1} &= \hat{P}_m y_m \\ \hat{y}_{m+2} &= (\hat{P}_m)^2 y_m \\ &\cdots \\ \hat{y}_{m+l} &= (\hat{P}_m)^l y_m \end{aligned} \tag{5}$$

where $\hat{P}_m$ is OLS (or MLE) estimates of VAR coefficients, i.e. $\hat{P}_m = [y_m' y_m]^{-1} y_m' y_{m+1}$.

2.3 OSF tests

The root mean square error (RMSE) of a forecast is defined as:

$$RMSE_j(l) = \sqrt{\frac{1}{T-l-m}\sum_{m=M}^{T-l}(y_{m+l} - \hat{y}_{j,m+l})^2} \tag{6}$$

where y_{m+l} is the true data, $\hat{y}_{j,m+l}$ is its out of sample forecasts from model j; M is the initial forecast origin. $l = 1,2,\ldots,h$ denotes the l -step ahead forecast. We look at the 4-quarter-ahead (4Q) and 8-quarter-ahead (8Q) forecasts. T is the sample size. $j = 1,2$ denotes the two competing models, say M1 is the DSGE model, M2 is the unrestricted VAR model. Then $RMSE_j(l)$ is the root mean squared forecast error for the l-step-ahead forecast of model j.

The OSF test is carried out on the ratio of the RMSE of the DSGE model to that of the VAR:

$$Ratio(l) = \frac{RMSE_{DSGE}(l)}{RMSE_{VAR}(l)} \tag{7}$$

Since it is hard to find the asymptotic distribution for the OSF Ratio test, we use Monte Carlo methods and when the error distribution is unknown, the bootstrap. By these methods, described in detail below, we obtain the empirical distribution of the OSF Ratio. From this distribution, we find (say) the 95% percentile and use it as the empirical critical value. Since the tests considered are one-sided tests, the p-value of the OSF Ratio test is the percentage of the empirical distribution above the test statistic. It should be noted that the empirical critical value varies with sample size, forecast origin and forecast horizons.

To compare the out-of-sample forecasting ability, there are two alternative statistics that focus on the difference of the minimum mean-squared forecast error (MSFE) between two nested models: the Diebold-Mariano and West (DMW) and the Clark-West (CW) statistics. Diebold and Mariano (1995) and West (1996) construct t-type statistics which are assumed to be asymptotically normal and where the sample difference between the two MSFE's are zero under the null. Clark and West (2006, 2007) provide an alternative DMW statistic that adjusts for the negative bias in the difference between the two MSFEs.

However in empirical analysis, both the DMW and CW test statistics take their critical values from their asymptotic distributions. Rogoff and Stavrakeva (2008) criticize the asymptotic CW test as oversized; an oversized asymptotic CW test would cause too many rejections of the null hypothesis. Rogoff and Stavrakeva (2008) and Onur Ince (2014) propose to use the bootstrapped OSF test to avoid this size distortion in small samples.

Our bootstrapped OSF test statistics are similar to these. There is not too much difference between the simulated asymptotic distributions of the RMSE ratio and the RMSE difference. But we focus on the ratio of the RMSEs between the DSGE and the VAR model, as this is the measure usually adopted in macroeconomic forecasting studies, such as those discussed here.

3. The power of OSF tests

3.1 Monte Carlo experiments

We follow the basic procedures of Le et al (2011) to design the Monte Carlo experiment. We take the model of Smets and Wouters (2007) for the US and adopt

their posterior modes for all parameters, including for error processes; the innovations are given their posterior standard errors with the normal distribution (Table 1A&1B, SW (2007)).

We set the sample size (T) at 200, and generate 1000 samples. We set the initial forecast origin (M) at 133. The VAR and DGSE autoregressive processes are initially estimated over the first 133 periods. The models were then used to forecast the data series 4- or 8-periods-ahead over the remaining 67 periods, with re-estimation every period (quarter). We find the distribution of this for the relevant null hypothesis under our small sample from our 1000 Monte Carlo samples. Our null hypothesis for the OSF tests is 1) the True DSGE model and 2) (discussed in section 4) the False DSGE model that marginally succeeds in improving the forecast.

We follow Le et al (2011) in specifying a False DSGE model. A False DSGE model is chosen by changing the parameters (A_0, A_1, B_0) in the true model by + or - $q\%$ alternately where q is the degree of falseness. We then extract the model residuals (z_t) from the data, re-estimate the error process and get $\hat{R}$. Le et al (2011) consider two ways to extract the model residuals (the Limited Information estimation method, LIML, which projects expectations by Instrumental Variables and the Exact Method, which projects them as the DSGE model solution) and find their differences are trivial. We use the Exact Method to estimate the model residuals and get $\hat{R}$ [2] Denoting the false parameters as $\Omega^F = \{A_0^F, A_1^F, B_0^F, \hat{R}\}$, we can derive A^F from Dynare as before. The OSFs are calculated as in (3), except that we use A^F rather than A. The RMSE of the False DSGE model is:

$$RMSE_{DSGE}^{F}(l) = \sqrt{\frac{1}{T-l-m}\sum_{m=M}^{T-l}(y_{m+l} - \hat{y}_{DSGE,m+l}^{F})^2} \qquad (8)$$

where $\hat{y}_{DSGE,m+l}^{F}$ is the OSF from the False DSGE model. The RMSE of the VAR model remains the same. Then we can obtain the ratio test statistic for each sample.

[2] We only reestimate the errors for a given False model (for each overlapping sample). If we reestimated the whole False model each period, it would have variable falseness.

$$Ratio(l) = \frac{RMSE^{F}_{DSGE}(l)}{RMSE_{VAR}(l)} \tag{9}$$

The power of the test is the probability of rejecting a hypothesis when it is false. In our OSF test, the power of the ratio test is the probability that the Ratio > the 5% critical value for the True distribution.

3.2 Asymptotic versus small sample distributions

We begin with a discussion of how the distribution for our typical 200-size sample differs from the asymptotic. In the absence of an analytical expression for the asymptotic distribution we use a sample of 1000 as a proxy (as can be seen from Figure 2 it is close to the t_{∞} distribution)- we raise both the sample used to obtain the forecasts and the subsequent sample used to make the forecasts, in proportion, i.e. by 5 times. In this way we obtain five times the size of sample for estimation and five times as many forecasts for the evaluation; this mimics the idea of raising the data available to 'very large' amounts. Figures 1 show that the 5% critical value differs by more than 10% between the two for the case shown here of the 4Q forecast which is typical.

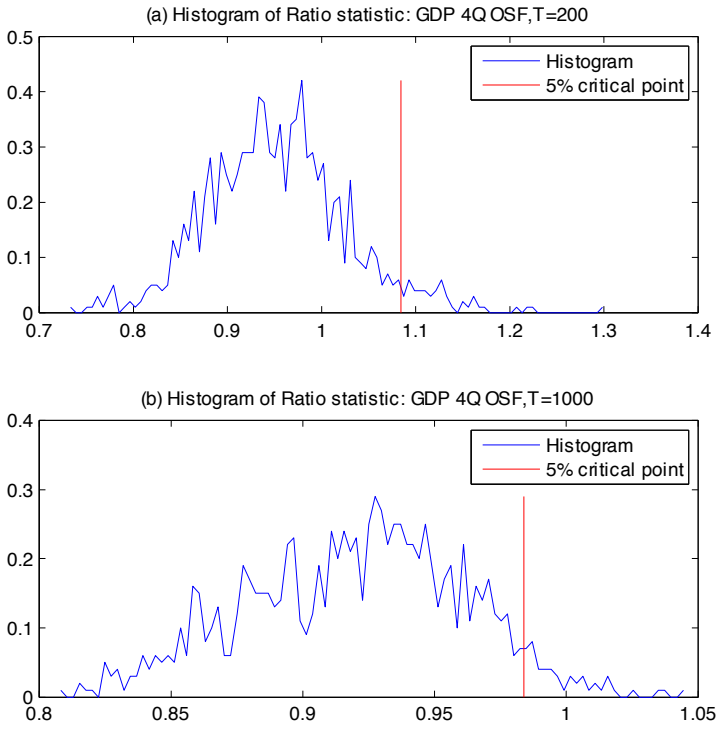

Figure 1: Asymptotic versus small sample distributions

We then normalise the ratio statistics by adjusting its mean and standard deviation. This is plotted agaist a normal distribution in figure 2. It can be observed that the large sample distribution is very close to a normal distribution. The 5% critical value for the normalized large sample ratio is 1.543, which is close to 5% critical value from the standard normal distribution (1.645).

In what follows all the distributions are based on Monte Carlo results for $T = 200$. For the sake of brevity we focus solely on the 5% confidence level test.

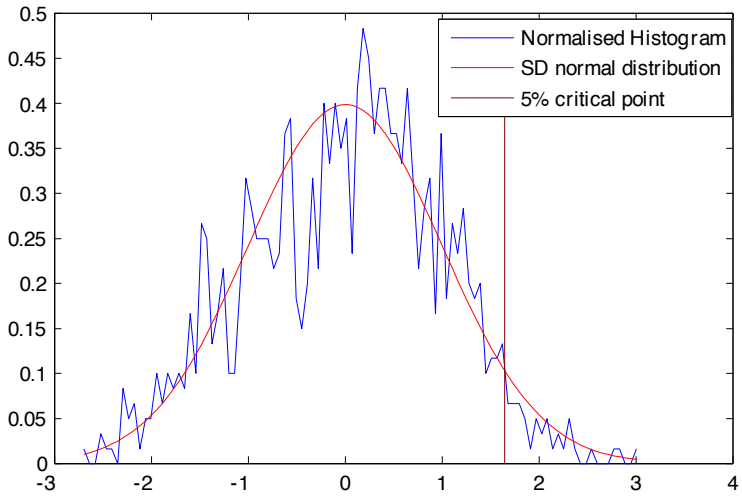

Figure 2: Normalized ratio statistics and standard normal distribution

3.3 Power of the specification test at 5% nominal value

The Power of the OSF tests at a 5% nominal value are reported in table 1. The first three sets of results are for each variable viewed alone. The last set relates to the joint forecast performance; for this we use the square root of the determinant of the joint forecast-error-covariance matrix (also used to measure the joint error in SW 2007)[3]. See appendix for the small sample distribution and the 5% critical value associated with the OSF tests in table 1.

[3]It is defined as follows. Let f_y, f_π, f_r be the OSF errors of output growth, inflation and interest rate respectively. Denote $f = (f_y, f\ , f_r)'$. Then f is a $(T - l - m) * 3$ matrix. We can calculate the covariance of f. The joint RMSE is defined as $\sqrt{|\operatorname{cov}(f)|}$.

Table 1: Power of OSF test

GDP growth			Inflation			Interest rate			Joint 3		
% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q
True	5.0	5.0	True	5.0	5.0	True	5.0	5.0	True	5.0	5.0
1	10.2	5.0	1	5.8	4.7	1	4.7	4.8	1	6.0	4.9
3	23.2	5.0	3	7.9	4.8	3	6.5	4.2	3	9.4	5.2
5	34.9	5.2	5	13.4	5.1	5	11.5	4.2	5	15.3	6.0
7	42.5	5.1	7	21.3	6.9	7	18.9	5.4	7	22.9	6.6
10	52.3	5.5	10	35.6	10.7	10	30.3	6.5	10	36.2	9.8
15	58.0	11.0	15	62.7	23.7	15	48.9	11.9	15	73.8	29.5
20	49.9	60.5	20	97.8	72.4	20	62.7	21.3	20	99.8	90.7

Notes on results in Table 1: (1) The 4Q-ahead GDP growth forecast is rejected less when the model is 20% False than when 15% False; this could arise from the reestimation of the model error processes that takes place when each model version is created; this reestimation can offset the effects of falseness of parameters. Thus in the 20% False model this offset could by chance be greater than for the 15%. (2) Sometimes the rejection rate for 95% confidence dips below 5%; this can happen for the same reason that error reestimation can offset the effect of parameter falseness. (3) The Joint 3 rejection rate cannot be obtained as the average of the three individual rejection rates because the forecast behaviour of the three variables may be correlated; thus if a forecast fails on one variable it is more likely to fail on another, raising the joint failure rate.

These results are obtained with stationary errors and with a VAR(1) as the benchmark model. We redid the analysis under the assumption that productivity was non-stationary. The results were very similar to those above. We further looked at a case of much lower forecastability, where we reduced the AR parameters of the error processes to a minimal 0.05 (on the grounds that persistence in data can be exploited by forecasters). Again the results were very similar, perhaps surprisingly. It seems that while absolute forecasting ability of a model, whether it is a DSGE or a VAR, is indeed reduced by lesser forecastability, relative forecasting ability is rather robust to data forecastability. Finally, we redid the original analysis using a VAR(2) as the benchmark; this also produced similar results to those above. All these variants, designed to check the robustness of our results, are to be found in Appendix 2.

What we see from Table 1 is that the power is weak. On a 1-year-ahead forecast, 4Q, the rejection rate of the DSGE model on its joint performance remains low at the one year horizon until the model reaches 20% falseness, and at the two year horizon does not get above 40% even when the model is 20% false. Notice also that the individual variable tests show some instability, which is due to the way the OSF uses reestimated error processes for each overlapping-sample forward projection: each time the errors are reestimated the full model in effect is changed and sometimes this

improves its forecasting performance, sometimes worsens it. Thus forecast performance does not always deteriorate with rising parameter falseness, When all variables are considered jointly this is much less of a problem as across the different variables the effects of reestimation on forecast performance are hardly correlated.

To put this RMSE test in perspective consider the power of the indirect inference Wald test, in sample using a VAR(1) on the same three variables (GDP, inflation and interest rates)- taken from Le et al (2012a) which also describes in full the procedures for obtaining the test, based on checking how far the DSGE model can generate in simulated samples the features found in the actual data sample.

Table 2: Rejection Rates for Wald and Likelihood Ratio for 3 Variable VAR(1)

% F	Wald in-sample II	Joint 3:4Q	:8Q
True	5.0	5.0	5.0
1	19.8	6.0	4.9
3	52.1	9.4	5.2
5	87.3	15.3	6.0
7	99.4	22.9	6.6
10	100.0	36.2	9.8
15	100.0	73.8	29.5
20	100.0	99.8	90.7

We see that the in-sample Wald II test has far more power. Why may this be the case? In forecasting, as we have just emphasised, DSGE models use fitted errors and when the model is mis-specified this creates larger errors which absorb the model's mis-specification; these new errors are projected into the future and could to some degree compensate for the poorer performance by the mis-specified parameters. To put this another way, as the DSGE model produces larger errors, reducing the relative input from the structural model proper, these larger errors take on some of the character of an unrestricted VAR. By contrast in indirect inference false errors compound the model's inability to generate the same data features as the actual data.

3.4 The connection between mis-specification and forecast improvement

For our small samples here we find that the cross-over point at which the DSGE model forecasts 1 year ahead less well on average than the unrestricted VAR is for output growth 1% false, for inflation and interest rates 7% false; for the three variables together it is also 7%. This reveals that the lower the power of the forecasting test for a variable the more useful are False models in improving

unrestricted VAR forecasts. Thus for output growth where power is higher, the DSGE model needs to be less than 1% false to improve the forecast; yet for inflation and interest rates where the power is very weak a model needs only to be less than 7% false to improve the forecast. This is illustrated in the two cases shown in Figure 3. In the lower one the false distribution with a mean RMSE ratio of unity (where the DSGE model is on average only as accurate as the unrestricted VAR) is 7% false; hence any model less false than this will have a distribution with a mean ratio of less than unity- and will therefore on average improve the forecast. In the upper one the false distribution with a mean RMSE ratio of unity is only 1% false; so to improve output growth forecasts you need a model that is less than 1% false. Essentially what is happening with weak power is that as the model becomes more false its RMSE ratio distribution moves little to the right, with the OSF performance deteriorating little; this, as we have pointed out, may be because as the model parameters worsen, the error parameters offset some of this worsening.

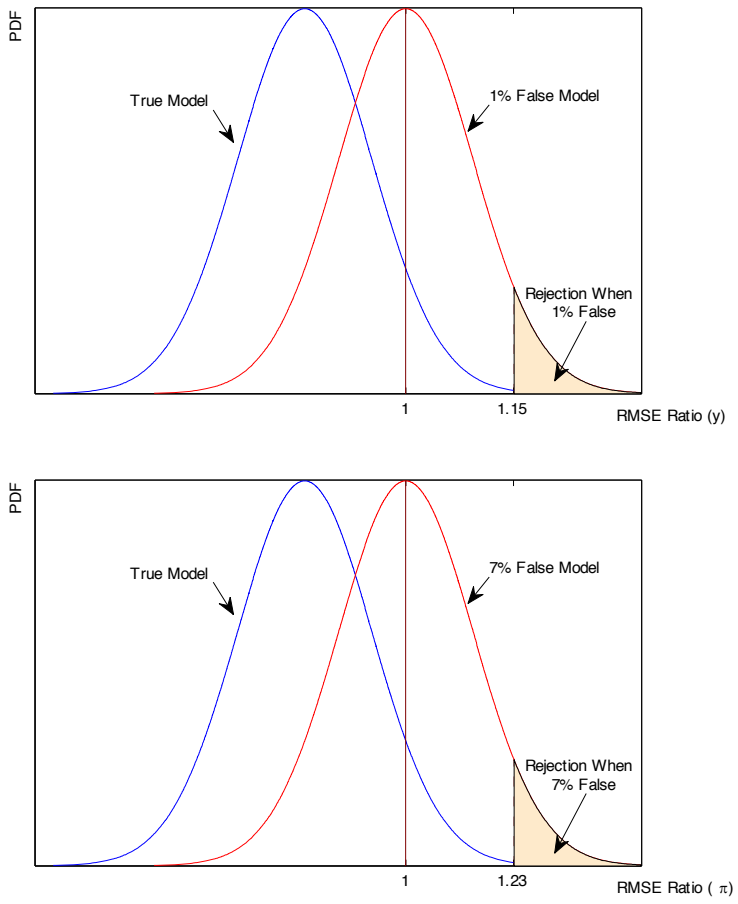

Figure 3: The connection between mis-specification and forecast improvement

What this shows is that if all a policymaker cares about is improving forecasts and the power of the forecast test is weak, then a poorly specified model may still suffice for improvement and will be worth using. This could well account for the willingness of central banks to use DSGE models in forecasting in spite of the evidence from other tests that they are mis-specified and so unreliable for policymaking. We now turn to how central banks can check on the forecasting capacity of their DSGE models using OSF tests.

4. OSF tests of whether a DSGE model improves forecasts

We now consider how policymakers could assure themselves of the forecasting capacity of their DSGE model. Here they set up the marginal forecast-failure model as the null hypothesis, illustrated as the red distributions in Figure 3. This is the structure of the Diebold-Mariano (1995) test widely used to test the forecast accuracy of models. Notice that policymakers can either look at the right hand tail, which tests the null against the alternative that the model forecasts worse; if they use this test they are assuming in the event of non-rejection that the model forecasts just better- the benefit of the doubt goes to the model. Or they can look at the left hand tail which tests against the alternative that the model forecasts better; if they use this test they are assuming in the event of non-rejection that the model is not worth using- the benefit of the doubt goes to the VAR forecast. If they obtain a result in the left hand tail, then they can be sure, at least with 95% confidence, that the model will improve forecasts. If they obtain a result in the right hand tail, then again they can be sure, at lest with 95% confidence, that the model will worsen forecasts. We need to check the power of each tail: how fast rejection rises on the RH tail as models get worse and on the LH tails how fast it rises as models get better. The situation is illustrated in figure 4.

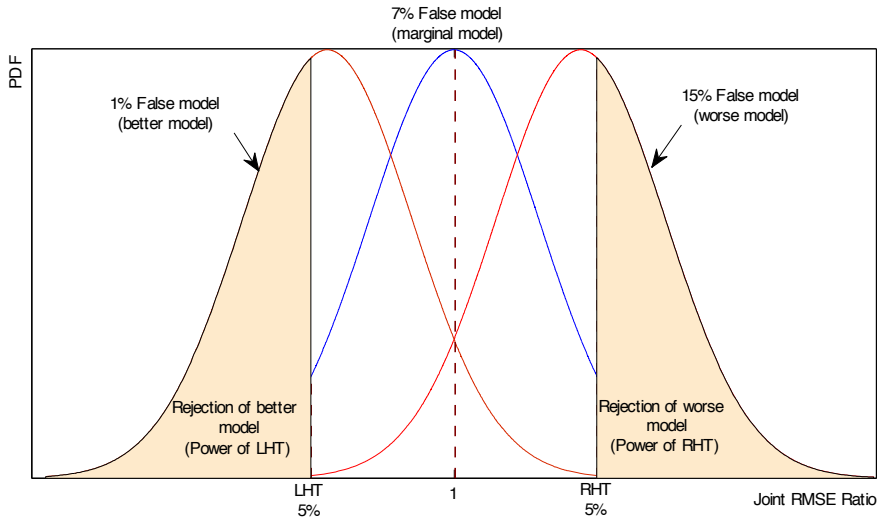

Figure 4: Illustration of LH and RH tails

4.1 Power of Left Hand and Right Hand tails

Table tests shows for the joint-3 case (the results for individual variables are reported in the appendix) the power of the Left Hand and Right Hand tails as just discussed. Thus for the LH tail we show the chances of less False models being rejected, while for the RH tail we show the chances of more False models being rejected.

The main problem with these tests remains that of poor power.

On the one hand, policymakers could use a DSGE model that was poor at forecasting without detection by the RH tail test. Thus for example a model that was 3% more false than the marginal one would only be rejected on the crucial 4Q-ahead test 11.3% of the time on the RH tail.

On the other hand, they could refuse to use a DSGE model that was good at forecasting without detection; for example a model that was 3% less False than the marginal one would only be rejected on the 4Q-ahead test by the LH tail 9.8% of the time.

We can design a more powerful test by going back to Table 2 and using simply the right hand tail as a test of specification. What is needed is a test of the DSGE model's specification (as true) that has power against a model that is so badly specified that it would marginally worsen forecasting performance on the joint 3 variables- the marginal forecast-failure model: as we have seen such a model is at the 4Q horizon 7% false and at the 8Q horizon 15% false. Now the power of OSF specification tests

against such a bad model is larger: Table 3 below shows that if on an OSF 4Q test at 95% confidence a model is not rejected (as true), then the marginal forecast-failure model (the 7% false model) has a 22.9% chance of rejection. On an 8Q test the equivalent model (15% false) has a 29.5% chance of rejection. Thus the OSF test has better power against the marginal forecast-failure model; but it is still quite weak.

Table 3: Power of OSF tests: LHT and RHT

Joint (Det)- RHTail			Joint (Det) -LHTail		
% F	4Q	8Q	% F	4Q	8Q
True			True	16.7	18.8
1			1	14.2	17.4
3			3	9.8	14.8
5			5	7.2	12.9
7	5.0		7	5.0	11.3
10	11.3		10		9.4
15	46.8	5.0	15		5.0
20	99.5	70.5	20		
25	100	100	25		
30	100	100	30		
35	100	100	35		
40	100	100	40		

Policymakers could however use the II in-sample test of whether the model is true also shown in that Table. Against the 4Q 7% false model it has power of 99.4%, and against the 8Q 15% false model power of 100%. Thus if policymakers could find a DSGE model that was not rejected by the II test, then they could have complete confidence that it could not worsen forecasts.

If no DSGE model can be found that fails to be rejected, then this strategy would not work and one must use the Diebold-Mariano test faute de mieux, on whatever DSGE model comes closest to passing the II specification test.

4.2 Reviewing the evidence of OSF tests

In this subsection we review some of the available OSF tests of DSGE models against time-series alternatives and see how we could interpret them in the light of these Monte Carlo experiments. Our aim is not to go through all such tests but merely to illustrate from some prominent ones how one might interpret the available evidence; we choose in particular those of SW(2007) and Gürkaynak et al (2013) for the SW (2007) model of the US on which our Monte Carlo experiment is also focused.

Table 4: DSGE/Time-series RMSE ratio for SW real-time data.

	RMSE:	4Q	8Q	4Q	8Q
Gürkaynak et al (2013)	VAR			RW	
	π	0.92	0.73	1.20	1.19
	Δy	0.68	0.63	0.70	0.69
	R	0.99	0.89	1.02	0.99
SW (2007)	VAR				
	π	0.54	0.32		
	y	0.80	0.77		
	R	0.98	0.72		
	$Joint$	0.80	0.66		

Source: Gurkaynak et al (2013), SW post-war model- for 1992-2007 as OSF period. NB they report the inverse of these ratios. SW(2007),SW model- for 1990-2004 as OSF period. NB they report the percentage gains relative to VAR(1) model; we convert these to RMSE ratios.

If we first consider the forecasting performance of these DSGE models, what we see from this summary table is that the RMSE ratio of DSGE models relative to different time-series forecasting methods varies from better to worse according to which variable and which time-series benchmark is considered: Gürkaynak et al (2013) note that there is a wide variety of relative RMSE performance. Wickens (2014) who reviews a wide range of country/variable forecasts finds the same. No joint performance measures are reported in these papers; however SW (2007)'s joint ratio comes out at 0.8 against a VAR(1) 4Q-ahead and 0.66 8Q-ahead.[4] Thus on these joint ratios the LH tail rejects the marginal forecast-failure model, strong evidence that the SW model forecasts better than a VAR1.

If we turn now to consider DSGE models' specification from these results, we see first that in general they do not reject these DSGE models. But because of the low power of the OSF tests, the same would be true with rather high probability of quite false models. Le et al (2011) show that the SW model is strongly rejected by the II Wald test, which is consistent with these OSF results, since as we have seen a false DSGE model may still forecast better than a VAR. They went on to find a version of the model, allowing for the existence of a competitive sector, that was not rejected for the Great Moderation period. By the arguments of this paper this model must also improve on time-series forecasts.

[4]SW (2007) calculate the overall percentage gain as $(\log(|cov(f_{VAR})|) - \log(|cov(f_{DGE})|)/2k$, where k is the number of variables (here=3). We convert this to joint ratio as follows: $(\log(|cov(f_{VAR})|) - \log(|cov(f_{DGE})|)/2k = -(log\sqrt{|cov(f_{DSG})|} - log\sqrt{|cov(f_{var})|})/k \approx -\frac{\sqrt{|cov(f_{DSG})|}-\sqrt{|cov(f_{VAR})|}}{\sqrt{|cov(f_{VAR})|}*k} = -\frac{JRMSE_{DSG}-JRMSE_{VAR}}{JRMSE_{VAR}*k} -(JointRatio+1)/k.$

5. Conclusions

OSF tests are now regularly carried out on DSGE models against time-series benchmarks such as the VAR1 used here as typical. These tests aim to discover how good DSGE models are in terms of a) specification b) forecasting performance. Our aim in this paper has been to discover how well these tests achieve these aims.

We have carried out a Monte Carlo experiment on a DSGE model of the type commonly used in central banks for forecasting purposes and on which out-of-sample (OSF) tests have been conducted. In this experiment we generated the small sample distribution of these tests and also their power as a test of specification; we found that the power of the tests for this purpose was extremely low. Thus when we apply these results to the reported tests of existing DSGE models we find that none of them are rejected on a 5% test; but the lack of power means that models that were substantially false would have a very high chance also of not being rejected. Researchers could therefore have little confidence in these tests for this purpose. We show that they would be better off using an in-sample indirect inference test of specification which has substantial power.

The reason for this relative weakness of OSF tests on DSGE models may be that the model errors, which are increased by the model mis-specification, nevertheless when projected forward compensate for the poorer forecast of the structural parameters. It follows that weak power implies that a DSGE model may be badly mis-specified and yet still forecast well. Thus a corollary of the low power is that DSGE models can still improve forecasts even when badly misspecified.

Viewed as tests of forecasting performance against the null of doing exactly as well as the VAR benchmark, OSF tests of DSGE models are used widely, with both the left hand tail of the distribution testing for significantly better performance and the right hand tail for significantly worse performance. Power is again rather weak, particularly on the left hand tail. An alternative would again be to use an in-sample indirect inference test of specification; if a DSGE model specification can be found that passes such a test, then it may not only be fit for policy analysis but will also almost definitely improve VAR forecasts.

References

[1] Adolfson, M., J. Linde, and M. Villani (2007), Forecasting Performance of an Open Economy Dynamic Stochastic General Equilibrium Model, Econometric Reviews 26(2-4), 289-328.

[2] Christoffel, K., G. Coenen, and A. Warne (2011), Forecasting with DSGE Models, in M. Clements and D. Hendry (eds), Oxford Handbook of Economic Forecasting, Oxford University Press, Oxford.

[3] Clark, T., and K. D. West (2006), Using Out-of-sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis, Journal of Econometrics 135, 155-186.

[4] Clark, T., and K. D. West (2007), Approximately Normal Tests for Equal Predictive Accuracy in Nested Models, Journal of Econometrics 138, 291–311.

[5] Clements, M., and D. Hendry, 2005, Evaluating a Model by Forecast Performance, Oxford Bulletin of Economics and Statistics 67 (Supplement), 931-956.

[6] Del Negro, M. and F. Schorfheide (2012), Forecasting with DSGE Models: Theory and Practice, in: G. Elliott and A. Timmermann (eds.), Handbook of Forecasting, Vol. 2, Elsevier.

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

[8] Edge, R.M. and R.S. Gürkaynak (2010), How Useful Are Estimated DSGE Model Forecasts for Central Bankers? Brookings Papers on Economic Activity 41(2), 209-259.

[9] Edge, R.M., M.T. Kiley and J.P. Laforte (2010), A Comparison of Forecast Performance Between Federal Reserve Forecasts, Simple Reduced-form Models, and a DSGE Model, Journal of Applied Econometrics 25(4), 720-754.

[10] Giacomini, R. and B. Rossi (2010), Forecast Comparisons in Unstable Environments, Journal of Applied Econometrics 25(4), 595-620.

[11] Gürkaynak, R.S., Kisacikoglu, B. and Rossi, B. (2013), Do DSGE models forecast more accurately out-of-sample than VAR models? CEPR discussion paper no. 9576, July 2013, CEPR, London.

[12] Ince, Onur, 2014, Forecasting exchange rates out-of-sample with panel methods and real-time data, Journal of International Money and Finance 43(C), 1-18.

[13] Juillard, M. (2001), DYNARE: a program for the simulation of rational expectations models. Computing in economics and finance 213. Society for Computational Economics.

[14] Le, V.P.M., D. Meenagh, P. Minford, M. Wickens (2011), How much nominal rigidity is there in the US economy --- testing a New Keynesian model using indirect inference. Journal of Economic Dynamics and Control 35(12), 2078-2104.

[15] Le, V.P.M., D. Meenagh, P. Minford, M. Wickens (2012a), Testing DSGE models by indirect inference and other methods: some Monte Carlo experiments. Cardiff Economics Working Paper E2012/15.

[16] Le, V.P.M., D. Meenagh, P. Minford, M. Wickens (2012b), What causes banking crises? An empirical investigation. Cardiff Economics Working Paper E2012/14.

[17] Rogoff, K.S. and V. Stavrakeva (2008), The Continuing Puzzle of Short-Horizon Exchange Rate Forecasting, NBER W.P. 14071.

[18] Smets, F. and R. Wouters (2007), Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach, American Economic Review, 97(3), 586-606.

[19] West, K.D. (1996), Asymptotic Inference about Predictive Ability, Econometrica 64,1067-1084

[20] Wickens, M. (2014), How Useful are DSGE Macroeconomic Models for Forecasting? Open Economies Review, 25(1), 171-193.

Appendix

Appendix 1: Small sample distribution and 5% critical values of OSF tests

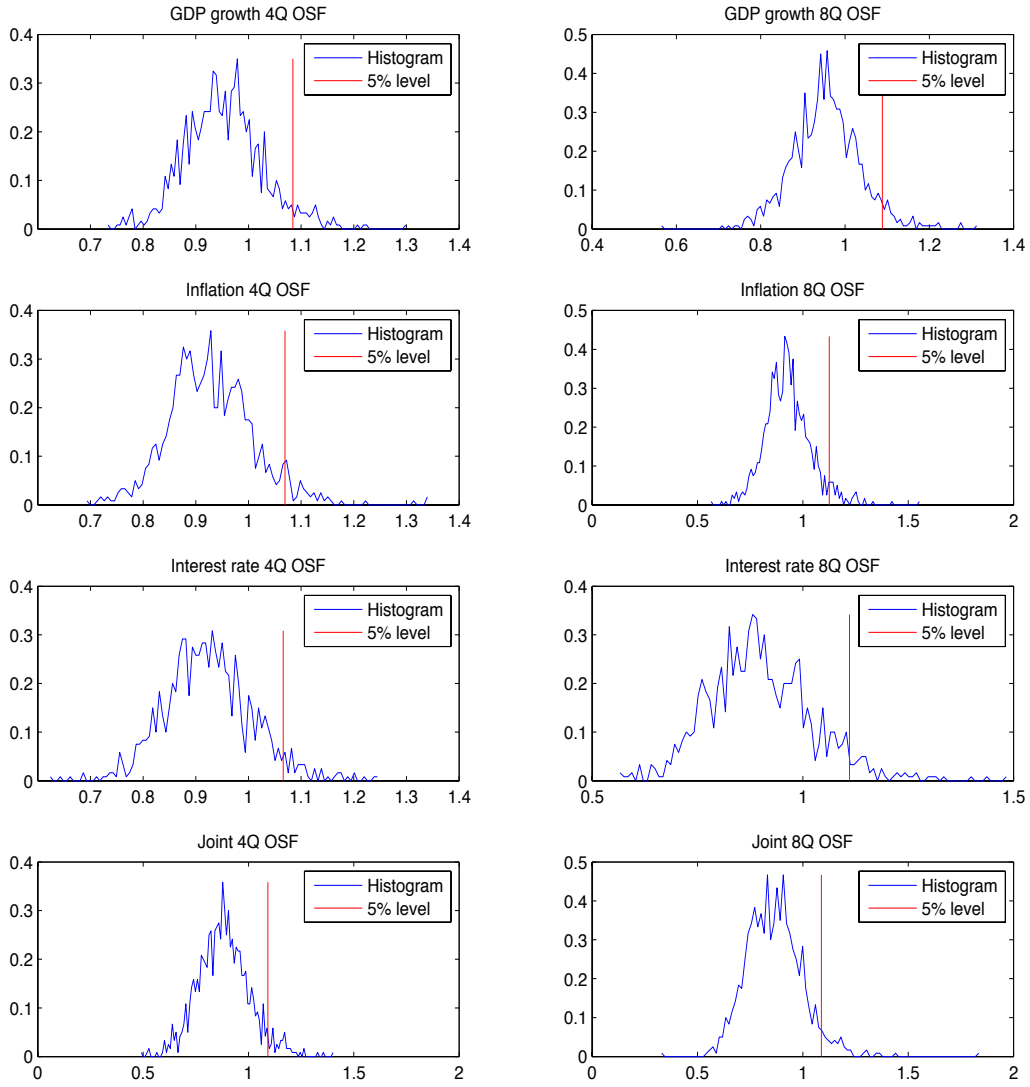

Figure 5: Historical distribution of ratio statistics: T=200

Table5: Empirical critical value at 5 percent level

	4Q	8Q
GDP growth	1.0844	1.0889
Inflation	1.0693	1.1257
Interest rate	1.0662	1.1107
Joint 3 variables	1.0922	1.0879

Appendix 2: Experiments with alternative error processes

a) productivity shock follows an I(1) process

We look here at the effect of non-stationarity in the shocks as exemplified by a non-stationary productivity process. We do not alter the status of other shocks because they are are typically found to be stationary for the SW model: for example in related work on the SW data Le et al (2012b) found that only productivity was non-stationary- see their Table 2 on p. 11.

Table 6: Power of OSF test

GDP growth			Inflation			Interest rate			Joint3 variables		
% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q
True	5.0	5.0	True	5.0	5.0	True	5.0	5.0	True	5.0	5.0
1	10.4	5.3	1	5.9	5.1	1	4.8	5.3	1	6.6	5.2
3	21.5	5.8	3	8.5	5.7	3	5.9	5.2	3	10.1	5.4
5	31.9	5.9	5	14.6	6.8	5	10.7	4.8	5	12.8	5.2
7	39.6	5.8	7	21.2	7.5	7	16.9	5.5	7	13.6	5.0
10	47.2	6.6	10	35.4	11.2	10	28.3	7.1	10	13.7	6.2
15	52.1	12.4	15	62.8	24.7	15	43.4	12.7	15	18.7	10.0
20	44.0	58.5	20	97.5	72.2	20	57.5	22.3	20	69.6	38.2

There is essentially no difference in the power of the test as productivity becomes I(1), thereby also making output I(1) (though leaving inflation and interest rates stationary). The change makes output growth positively instead of negatively autocorrelated and so may well make little difference to how easy it is to forecast.

The choice on stationarity is dictated by the general absence of unit roots in shocks other than productivity- for example in related work on the SW data Le et al (2012) found that only productivity was non-stationary- see Table 2 on p. 11 of "What causes banking crises? An empirical investigation" by Vo Phuong Mai Le, David Meenagh and Patrick Minford, Working Paper No. E2012/14, Cardiff University, Economics Section, Cardiff Business School, June 2012, updated April 2013- available from Minford repec page.

b) altering the forecastability of the economy

One might think that the power of the test would be affected by ease of forecasting the economy. We look at this issue by reducing the AR coefficients of the error processes to 0.05 from their SW values.

Table 7: Power of OSF test

GDP growth			Inflation			Interest rate			Joint3 variables		
% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q
True	5.0	5.0	True	5.0	5.0	True	5.0	5.0	True	5.0	5.0
1	5.4	3.7	1	5.3	5.9	1	4.2	3.8	1	5.1	5.0
3	5.6	3.3	3	6.3	8.7	3	3.4	2.8	3	6.3	6.6
5	5.4	3.6	5	8.9	11.2	5	5.4	3.3	5	10.0	10.1
7	5.1	5.9	7	14.9	16.1	7	8.0	3.9	7	17.4	15.8
10	4.8	14.8	10	31.8	31.0	10	13.6	6.3	10	37.0	31.9
15	5.4	46.0	15	88.6	73.0	15	30.2	20.3	15	88.0	76.6
20	10.2	93.3	20	100	100	20	56.7	50.6	20	100	100

What we see the power that is not dissimilar to that in our original Table.

c) altering the benchmark model

One might be concerned that the power of the test would be affected by using high order VARs. So we choose VAR(2) as benchmark model and redo the power of the test. The results are reported in the table below.

Table 8: Power of OSF test

GDP growth			Inflation			Interest rate			Joint3 variables		
% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q
True	5.0	5.0	True	5.0	5.0	True	5.0	5.0	True	5.0	5.0
1	5.7	4.5	1	5.1	5.1	1	4.9	4.7	1	5.6	5.2
3	9.7	4.2	3	5.6	5.1	3	5.7	4.5	3	5.6	5.3
5	14.8	4.4	5	6.9	5.8	5	7.4	4.5	5	6.4	5.4
7	18.2	4.8	7	8.5	6.1	7	9.9	5.1	7	7.5	5.2
10	22.7	5.2	10	13.1	8.0	10	12.1	5.5	10	10.6	6.4
15	24.7	7.5	15	27.9	13.9	15	16.2	8.1	15	24.7	8.7
20	20.5	38.5	20	69.0	45.3	20	22.2	12.6	20	87.0	42.5

With VAR(2) as the benchmark model, the OSF tests have similarly low power. The AR(2) coefficients are mostly insignificant; including high order terms worsens the VAR's forecast capacity. This is also consistent with other literature (e.g. SW 2007, Wickens 2014) in which a VAR(1) is often chosen as the benchmark model.

Appendix 3: OSF tests of whether a DSGE model improves forecasts for individual variables

Table 9: Power of OSF test: RHL

GDP growth			Inflation			Interest rate			Joint (Det)		
% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q
True			True			True			True		
1	5.0		1			1			1		
3	14.6		3			3			3		
5	22.7		5			5			5		
7	29.7		7	5.0		7	5.0		7	5.0	
10	38.5		10	12.3	5.0	10	12.9		10	11.3	
15	44.1	5.0	15	38.8	13.1	15	26.3		15	46.8	5.0
20	32.5	49.2	20	91.4	60.3	20	39.9	5.0	20	99.5	70.5
25	100	100	25	100	100	25	60.9	12.8	25	100	100
30	100	100	30	100	100	30	65.7	15.4	30	100	100
35	100	100	35	100	100	35	71.8	20.4	35	100	100
40	100	100	40	100	100	40	76.6	26.7	40	100	100

Table 10: Power of OSF test: LHT

GDP growth			Inflation			Interest rate			Joint (Det)		
% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q	% F	4Q	8Q
True	6.3	8.7	True	10.9	8.5	True	14.0	20.5	True	16.7	18.8
1	5.0	7.2	1	9.8	8.3	1	11.5	20.4	1	14.2	17.4
3		6.6	3	7.1	7.7	3	8.5	18.7	3	9.8	14.8
5		6.1	5	5.7	6.7	5	6.3	16.2	5	7.2	12.9
7		5.7	7	5.0	5.6	7	5.0	14.3	7	5.0	11.3
10		5.3	10		5.0	10		10.9	10		9.4
15		5.0	15			15		7.4	15		5.0
20			20			20		5.0	20		
25			25			25			25		
30			30			30			30		
35			35			35			35		
40			40			40			40		