Cardiff Economics Working Papers

Working Paper No. E2015/2

Small sample performance of indirect inference on DSGE models

Vo Phuong Mai Le, David Meenagh, Patrick Minford and Michael Wickens

January 2015

Cardiff Business School
Aberconway Building
Colum Drive
Cardiff CF10 3EU
United Kingdom
t: +44 (0)29 2087 4000
f: +44 (0)29 2087 4419
business.cardiff.ac.uk

Small sample performance of indirect inference on DSGE models[*]

Vo Phuong Mai Le (Cardiff University)[†]
David Meenagh (Cardiff University)[‡]
Patrick Minford (Cardiff University and CEPR)[§]
Michael Wickens (Cardiff University, University of York and CEPR)[¶]

January 2015

Abstract

Using Monte Carlo experiments, we examine the performance of indirect inference tests of DSGE models in small samples, using various models in widespread use. We compare these with tests based on direct inference (using the Likelihood Ratio). We find that both tests have power so that a substantially false model will tend to be rejected by both; but that the power of the indirect inference test is by far the greater, necessitating re-estimation to ensure that the model is tested in its fullest sense. We also find that the small-sample bias with indirect estimation is around half of that with maximum likelihood estimation.

JEL Classification: C12, C32, C52, E1

Keywords: Bootstrap, DSGE, Indirect Inference, Likelihood Ratio, New Classical, New Keynesian, Wald statistic

[*] We thank participants in the 2014 Brunel Macro-Finance Conference for helpful comments on an earlier version. We are also grateful to High Performance Computing Wales and our HPC partner, OSTC, for access to the HPC super-computing network.

[†] LeVP@cf.ac.uk; Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

[‡] Meenaghd@cf.ac.uk; Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

[§] Patrick.minford@btinternet.com; Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

[¶] mike.wickens@york.ac.uk; Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

1 Introduction

An unresolved issue in macroeconomics is how best to evaluate the empirical performance of DSGE models, especially those estimated by Bayesian methods. In this paper we compare a relatively new type of test, indirect inference, with a standard procedure, the Likelihood Ratio test. Our main concern is the performance of these tests in small samples, though we will refer to asymptotic properties where known. Our main finding is that the power of the likelihood ratio test is rather weak relative to that of the indirect inference test. We also consider why we find this.

The paper is set out as follows. In section 2 we consider how we might evaluate a DSGE model empirically. In section 3 we review the main features of the indirect inference testing procedure as implemented in this paper. In section 4 we compare the small sample properties of tests based on indirect inference with the Likelihood Ratio test that is used in direct inference. The comparison is based on Monte Carlo experiments on the widely used DSGE model introduced by Christiano, Eichenbaum and Evans (2005) and estimated by Smets and Wouters (2003, 2007) on EU and US data. Initially, we use stationary data. In section 5 we extend the analysis to non-stationary data and to the three-equation New Keynesian representation of the model of Clarida, Gali and Gertler (1999), again on both stationary and non-stationary data. In section 6 we consider why the two testing methods have such different power. Our final section presents our conclusions from these comparisons.

2 The empirical evaluation of DSGE models

DSGE models emerged largely as a response to the perceived shortcomings of previous formulations of macroeconometric models. The main complaints were that these macroeconometric models were not structural — despite being referred to as structural macroeconometric models — and so were subject to Lucas's critique that they could not be used for policy evaluation (Lucas, 1976), that they were not general equilibrium models of the economy but, rather, they comprised a set of partial equilibrium equations with no necessary coherent structure (for example, the Brookings model), that they incorporated 'incredible' identifying restrictions (Sims, 1980) and that they over-fitted the data through data-mining. For all their theoretical advantages, the strong simplifying restrictions on the structure of DSGE models resulted in a severe deterioration of fit compared to structural macroeconometric models with their ad hoc supply and demand functions, their flexible lagged adjustment mechanisms and their serially correlated structural errors.

There have been various reactions to the empirical failures of DSGE models. The early version of the DSGE model, the RBC model, was perceived to have four main faults: predicted consumption was too smooth compared with the data, real wages were too flexible resulting in employment being too stable, the predicted real interest rate was too closely related to output and the model, being real, could not admit real effects arising from nominal rigidities. In retrospect, however, this empirical examination was limited and flawed. Typically, the model was driven by a single real stochastic shock (to productivity); there were no nominal shocks; and the model's dynamic structure was derived solely from budget constraints and the capital accumulation equation. Subsequent developments of the DSGE model aimed to address these limitations, and other specification issues, and they had some empirical success. Nevertheless, even this success has been questioned; for example Le et al. (2011) reject the widely acclaimed model of Smets-Wouters (2007).

Another reaction, mainly from econometricians, is the criticism that DSGE models have been calibrated (to an economy) rather than estimated and tested using traditional methods, and when estimated and tested using classical econometric methods, such as the Likelihood Ratio test, they are usually found to perform poorly and are rejected. Sargent[1], discussing the response of Lucas and Prescott to these rejections, is quoted as saying that they thought

[1] In an interview Sargent remarked of the early days of testing DSGE models: '...my recollection is that Bob Lucas and Ed Prescott were initially very enthusiastic about rational expectations econometrics. After all, it simply involved imposing on ourselves the same high standards we had criticized the Keynesians for failing to live up to. But after about five years of doing likelihood ratio tests on rational expectations models, I recall Bob Lucas and Ed Prescott both telling me that those tests were rejecting too many good models.' Tom Sargent, interviewed by Evans and Honkapohja (2005).

that 'those tests were rejecting too many good models'.

Current practice is to try to get around this problem by estimating DSGE models using Bayesian rather than classical estimation methods. Compared with calibration, Bayesian methods allow some flexibility in the prior beliefs about the structural parameters and permit the data to affect the final estimates. Calibrated parameters or, equivalently, the priors used in Bayesian estimation, often come from other studies or from micro-data estimates. Hansen and Heckman (1996) point out that the justification for these is weak: other studies generally come up with a wide variety of estimates, while micro-estimates may well not survive aggregation. If the priors cannot be justified and uninformative priors are substituted, then Bayesian estimation simply amounts to classical ML in which case test statistics are usually based on the Likelihood Ratio. The frequency of rejection by such classical testing methods is an issue of concern in this paper.

A more radical reaction to the empirical failures of DSGE models has been to say that they are all misspecified and so should not be tested by the usual econometric methods which would always reject them — see Canova (1994). If all models are false, instead of testing them in the classical manner under the null hypothesis that they are true, one should use a descriptive statistic to assess the 'closeness' of the model to the data. Canova (1994), for example, remarks that one should ask "how true is your false model?" and assess this using a closeness measure. Various econometricians — for example Watson (1993), Canova (1994, 1995, 2005), Del Negro and Schorfheide (2004, 2006) — have shown an interest in evaluating DSGE models in this way.

We adopt a somewhat different approach that restores the role of formal statistical tests of DSGE models and echoes the widely accepted foundations of economic testing methodology laid down by Friedman (1953). Plainly no DSGE model, or indeed no model of any sort, can be literally true as the 'real world' is too complex to be represented by a model that is 'true' in this literal sense and the 'real world' is not a model. In this sense, therefore, all DSGE models are literally false or 'mis-specified'. Nevertheless an abstract model plus its implied residuals which represent other influences as exogenous error processes, may be able to mimic the data; if so, then according to usual econometric usage, the model would be 'well specified'. The criterion by which Friedman judged a theory was its potential explanatory power in relation to its simplicity. He gave the example of perfect competition which, although never actually existing, closely predicts the behaviour of industries with a high degree of competition. According to Friedman, a model should be tested, not for its 'literal truth', but 'as if it is true'. Thus, even though a macroeconomic model may be a gross simplification of a more complex reality, it should be tested on its ability to explain the data it was designed to account for by measuring the probability that the data could be generated by the model. In this spirit we assess a model using formal misspecifications tests. The probability of rejection gives a measure of the model's 'closeness' to the facts. This procedure can be extended to a sub-set of the variables of the model rather than all variables. In this way, it should be possible to isolate which features of the data the model is able to mimic; different models have different strengths and weaknesses ('horses for courses') and our procedure can tease these out of the tests.

The test criterion may be formulated in a number of ways. It could, for example, be interpreted as a comparison of the values of the likelihood function for the DSGE model, or of a model designed to represent the DSGE model (an auxiliary model), or it could be based on the mean square prediction error of the raw data or on the impulse response functions obtained from these models or, as explained in more detail later, it could be based on a comparison of the coefficients of the auxiliary model being associated with the DSGE model. These criteria fall into two main groups: on the one hand, closeness to raw data, size of mean squared errors and 'likelihood' and, on the other hand, closeness to data features, to stylised facts or to coefficients of VARs or VECMs. Within each of these two categories the criteria can be regarded as mapping into each other so that there are equivalences between them; for example, a VAR implies sets of moments/cross-moments and vice versa. We discuss both types in this paper; we treat the Likelihood Ratio as our representative of the first type and the coefficients of a VAR as our representative of the second.

Before DSGE models were proposed as an alternative to structural macroeconometric models, in response to the latter's failings, Sims (1980) suggested modelling the macroeconomy as a VAR. This is now widely used in macroeconometrics as a way of representing the data in a theory-free manner in order, for example, to estimate impulse response functions or for forecasting where they perform as well, or sometimes better, than structural models, including DSGE models, see Wieland and Wolters (2012) and Wickens (2014). Moreover, it can be shown that the solution to a (possibly linearized) DSGE model where the exogenous variables are generated by a VAR is, in general, a VAR with restrictions on its coefficients, Wickens (2014). It follows that a VAR is the natural auxiliary model to use for evaluating how closely a DSGE model fits the data whichever of the measures above are chosen for the comparison. The data can be represented by an unrestricted VAR and the DSGE model by the appropriately restricted VAR; the two sets of estimates can then be compared according to the chosen measure.

The apparent difficulty in implementing this procedure lies in estimating the restricted VAR. Indirect inference provides a simple solution. Having estimated the DSGE model by whatever means — the most widely used at present being Bayesian estimation — the model can be simulated to provide data consistent with the estimated model using the errors backed out of the model. The auxiliary model is then estimated unrestrictedly both on these simulated data and on the original data. The properties of the two sets of VAR estimates can then be compared using the chosen measure. More precise details of how we carry out this indirect inference procedure in this paper are given in the next section.

3 Model evaluation by indirect inference

Indirect inference provides a classical statistical inferential framework for judging a calibrated or already, but maybe partially, estimated model whilst maintaining the basic idea employed in the evaluation of the early RBC models of comparing the moments generated by data simulated from the model with actual data. An extension of this procedure is to posit a general but simple formal model (an auxiliary model) — in effect the conditional mean of the distribution of the data — and base the comparison on features of this model, estimated from simulated and actual data. If necessary these features can be supplemented with moments and other measures directly generated by the data and model simulations.

Indirect inference on structural models may be distinguished from indirect estimation of structural models. Indirect estimation has been widely used for some time, see Smith (1993), Gregory and Smith (1991,1993), Gourieroux et al. (1993), Gourieroux and Monfort (1995) and Canova (2005). In indirect estimation the parameters of the structural model are chosen so that when this model is simulated it generates estimates of the auxiliary model similar to those obtained from actual data. The optimal choice of parameters for the structural model are those that minimise the distance between the two sets of estimated coefficients of the auxiliary model. Common choices for the auxiliary model are the moments of the data, the score and a VAR. Indirect estimates are asymptotically normal and consistent, like ML. These properties do not depend on the precise nature of the auxiliary model provided the function to be tested is a unique mapping of the parameters of the auxiliary model. Clearly, the auxiliary model should also capture as closely as possible the data features of the DSGE model on the hypothesis that it is true.

Using indirect inference for model evaluation does not necessarily involve the estimation of the parameters of the structural model. These can be taken as given. They might be calibrated or obtained using Bayesian or some other form of estimation. If the structural model is correct then its predictions about the auxiliary model estimated from data simulated from the given structural model should match those based on actual data. These predictions relate to particular properties (functions of the parameters) of the auxiliary model such as its coefficients, its impulse response functions or just the data moments. A test of the structural model may be based on the significance of the difference between estimates of these functions derived from the two sets of data. On the null hypothesis that the structural model is 'true' there should be no significant difference. In carrying out this test, rather than rely on the asymptotic distribution of the test statistic, we estimate its small sample distribution and use this.

Our choice of auxiliary model exploits the fact that the solution to a log-linearised DSGE model can be represented as a restricted VARMA and also often by a VAR (or if not then closely represented by a VAR). For further discussion on the use of a VAR to represent a DSGE model, see for example Canova (2005), Dave and DeJong (2007), Del Negro and Schorfheide (2004, 2006) and Del Negro et al. (2007a,b) (together with the comments by Christiano (2007), Gallant (2007), Sims (2007), Faust (2007) and Kilian (2007)), and Fernandez-Villaverde et al. (2007). A levels VAR can be used if the shocks are stationary, but a VECM is required, as discussed below, if there are non-stationary shocks. The structural restrictions of the DSGE model are reflected in the data simulated from the model and will be consistent with a restricted version of the VAR[2]. The model can therefore be tested by comparing unrestricted VAR estimates (or some function of these estimates such as the value of the log-likelihood function or the impulse response functions) derived using data simulated from the DSGE model with unrestricted VAR estimates obtained from actual data.

The model evaluation criterion we use is based on the difference between the vector of relevant VAR coefficients from simulated and actual data as represented by a Wald statistic. If the DSGE model is correct (the null hypothesis) then the simulated data, and the VAR estimates based on these data, will not be significantly different from those derived from the actual data. The method is in essence extremely simple; although it is numerically taxing, with modern computer resources, it can be carried out quickly. The simulated data from the DSGE model are obtained by bootstrapping the model using the structural shocks implied

[2] This requires that the model is identified, as assumed here. Le, Minford and Wickens (2013) propose a numerical test for identification based on indirect inference and show that both the SW and the New Keynesian 3-equation models are identified according to it.

by the given (or previously estimated) model and computed from the historical data. The test then compares the VAR coefficients estimated on the actual data with the distribution of VAR coefficient estimates derived from multiple independent sets of the simulated data. We then use a Wald statistic (WS) based on the difference between a_T, the estimates of the VAR coefficients derived from actual data, and $\overline{a_S(\theta_0)}$, the mean of their distribution based on the simulated data, which is given by:

$$WS = (a_T - \overline{a_S(\theta_0)})' W(\theta_0)(a_T - \overline{a_S(\theta_0)})$$

where $W(\theta_0)$ is the inverse of the variance-covariance matrix of the distribution of simulated estimates a_S. and θ_0 is the vector of parameters of the DSGE model on the null hypothesis that it is true.

As previously noted, we are not compelled to use the VAR coefficients in this formula: thus one could use other data 'descriptors' considered to be key features of the data that the model should match — these could be particular impulse response functions (such as to a monetary policy shock) or particular moments (such as the correlations of various variables with output). However, such measures are functions of the VAR coefficients and it seems that a parsimonious set of features is these coefficients themselves. There are still issues about which variables to include in the VAR (or equivalently whether to focus only on a subset of VAR coefficients related to these variables) and what order of lags the VAR should be. Also it is usual to include the variances of the data or of the VAR residuals as a measure of the model's ability to match variation. We discuss these issues further below.

We can show where in the Wald statistic's bootstrap distribution the Wald statistic based on the data lies (the Wald percentile). We can also show the Mahalanobis Distance based on the same joint distribution, normalised as a t-statistic, and also the equivalent Wald p-value, as an overall measure of closeness between the model and the data.[3] In Le et al. (2011) we

[3] The Mahalanobis Distance is the square root of the Wald value. As the square root of a chi-squared distribution, it can be converted into a t-statistic by adjusting the mean and the size. We normalise this here by ensuring that the resulting t-statistic is 1.645 at the 95% point of the distribution.

applied this test to a well-known model of the US, that of Smets and Wouters (2007; qv). We found that the Bayesian estimates of the Smets and Wouters (SW) model were rejected for both the full post-war sample and for a more limited post-1984 (Great Moderation) sample. We then modified the model by adding competitive goods and labour market sectors. Using a powerful Simulated Annealing algorithm, we searched for values of the parameters of the modified model that might improve the Wald statistic and succeeded in finding such a set of parameters for the post-1984 sample.

A variety of practical issues concerning the use of the bootstrap and the robustness of these methods more generally are dealt with in Le at al. (2011). A particular concern with the bootstrap has been its consistency under conditions of near-unit roots. Several authors (e.g. Basawa et al., 1991, Hansen (1999) and Horowitz, 2001a,b) have noted that asymptotic distribution theory is unlikely to provide a good guide to the bootstrap distribution of the AR coefficient if the leading root of the process is a unit root or is close to a unit root. This is also likely to apply to the coefficients of a VAR when the leading root is close to unity and may therefore affect indirect inference where a VAR is used as the auxiliary model. In Le et al. (2011) we carried out a Monte Carlo experiment to check whether this was a problem in models such as the SW model. We found that the bootstrap was reasonably accurate in small samples, converged asymptotically on the appropriate chi-squared distribution and, being asymptotically chi-squared, satisfied the usual requirement for consistency of being asymptotically pivotal.

4 Comparing Indirect and Direct Inference testing methods

It is useful to consider how indirect inference is related to the familiar benchmark of direct inference. We focus on the Likelihood Ratio as representative of direct inference. We seek to compare the distribution of the Wald statistic for a test of certain features of the data with the corresponding distribution for likelihood ratio tests. We are particularly interested in the behaviour of these distributions on the null hypothesis and the power of the tests as the model deviates increasingly from its specification under the null hypothesis. We address these questions using Monte Carlo experiments.

4.1 Some preliminary experiments comparing indirect with direct inference

We base our comparison on tests of the performance of DSGE models. Our first comparison is based on the SW model of the US, estimated over the whole post-war sample ($1947Q1 - 2004Q4$), and with a VAR as the auxiliary model. We treat the SW model as true. The focus of the two tests is slightly different: direct inference asks how closely the model forecasts current data while indirect inference asks how closely the model replicates properties of the auxiliary model estimated from the data. For direct inference we use a likelihood ratio (LR) test of the DSGE model against the unrestricted VAR. In effect, this test shows how well the DSGE model forecasts the 'data' compared with an unrestricted VAR estimated on that data.

We examine the power of the Wald test by positing a variety of false models, increasing in their order of falseness. We generate the falseness by introducing a rising degree of numerical mis-specification for the model parameters. Thus we construct a False DSGE model whose parameters were moved $x\%$ away from their true values in both directions in an alternating manner (even-numbered parameters positive, odd ones negative); similarly, we alter the higher moments of the error processes (standard deviation, skewness and kurtosis) by the same $+/-x\%$. We may think of this False Model as having been proposed as potentially 'true' following previous calibration or estimation of the original model but which was at the time thought to be mis-specified.[4]

[4] The 'falseness' of the original model specification may arise due to the researcher not allowing the data to force the estimated parameters beyond some range that has been wrongly imposed by incorrect theoretical requirements placed on the model. If the researcher specifies a general model that nests the true model then

Many of the structural disturbances in the SW model are serially correlated, some very highly. These autocorrelated errors in a DSGE model are regarded as exogenous shocks (or combinations of shocks) to the model's specification, such as preferences, mark-ups, or technological change, the type of shock depending on which equation they appear in. Although they are, therefore, effectively the model's exogenous variables, they are not observable except as structural residuals in these equations. The significance of this is that, when the False models are constructed, the autocorrelation processes of the resulting structural errors are likely to be different. This difference is a marker of the model's mis-specification, as is the falseness of the structural coefficients. In order to give the model the best chance of not being rejected by the LR test, therefore, it is normal to re-estimate the autocorrelation processes of the structural errors. For the Wald test we falsify all model elements, structural and autocorrelation coefficients, and innovation properties, by the same $+/-x\%$.

In evaluating the power of the test based on indirect inference using our Monte Carlo procedure we generate 10,000 samples from some True model (where we take an error distribution with the variance, skewness and kurtosis found in the SW model errors), and find the distribution of the Wald for these True samples. We then generate a set of 10,000 samples from the False model with parameters θ and calculate the Wald distribution for this False Model. We then calculate how many of the actual samples from the True model would reject the False Model on this calculated distribution with 95% confidence. This gives us the rejection rate for a given percentage degree $+/-x\%$ of mis-specification, spread evenly across the elements of the model. We use 10,000 samples because the size of the variance-covariance matrix of the VAR coefficients is large for VARs with a large number of variables.[5]

estimation by indirect inference would necessarily converge on the parameter estimates that are not rejected by the tests. Accordingly tests would not reject this (well-specified) model. Thus the tests have power against estimated models that are mis-specified so that the true parameters cannot be recovered. Any estimation procedure that incorrectly imposes parameter values on a true model will generate such mis-specification.

In the case of the LR test the same argument applies, except that the estimator in this case FIML. Thus again the LR test cannot have power against a well-specified model that is freely estimated by FIML.

[5] We assume in this the accuracy of the bootstrap itself as an estimate of the distribution; the bootstrap substitutes repeated drawings from errors in a particular sample for repeated drawings from the underlying population.

Le et al. (2011) evaluate the accuracy of the bootstrap for the Wald distribution and find it to be fairly

In evaluating the power of the test under direct inference we need to determine how well the DSGE model forecasts the simulated data generated by the True Model compared with a VAR model fitted to these data. We use the first 1000 samples; no more are needed in this case. The DSGE model is given a parameter set θ and for each sample the residuals and their autoregressive parameters ρ are extracted by LIML (McCallum, 1976; Wickens, 1982). The IV procedure is implemented using the VAR to project the rational expectations in each structural equation; the residual is then backed out of the resulting equation. In the forecasting test the model is given at each stage together with the lagged data, including the lagged errors. We assume that since the lagged errors are observed in each simulated sample, the researcher can also estimate the implied ρs for the sample errors and use these in the forecast. We assume the researcher does this by LIML which is a robust method — clearly the DSGE model's forecasting capacity is helped by the presence of these autoregressive error processes. We find the distribution of the LR when θ is the true model. We then apply the 5% critical value from this to the False model LR value for each True sample and obtain the rejection rate for the False Model. Further False models are obtained by changing the parameters θ by + or $-x\%$.[6]

Table 1 shows that the power of the Indirect Inference Wald test is substantially greater

high.

[6] The two tests are compared for the same degree of falseness of the structural coefficients, with the error properties determined according to the each test's own logic. Thus for the Wald test, the error properties have the same degree of falseness as the structural coefficients so that overall model falseness is the same, rising steadily to give a smooth power function. For the LR test, the error properties are determined by reestimation, the normal test practice; the model's falseness rises smoothly with the falseness of the structural coefficients, and their accompanying implied error processes.

Were the LR error properties set at the same degree of falseness as for the Wald, the model's forecasting performance would go off track and the test would sharply reject, simply for this reason. Thus it would not be testing the model but arbitrarily false residuals — hence normal practice.

If, per contra, we were to reestimate the errors in the Wald test for conformity with the LR test, the falseness of the error properties would rise sharply due to estimation error, raising overall model falseness with it, so derailing the smooth rise in falseness for the power function.

To obtain exactly the same overall falseness of both tests, one needs to compare them with the same (true) error properties; this comparison is done in section 6, where it again shows much greater power from the Wald test. Of course in practice neither test would be appropriately carried out this way, nor could they since the tester is not told the true errors.

The comparisons of the two power functions as done here represents how rejection rates rise as these two different tests are applied in practice to models of smoothly increasing falseness.

than that of the Direct Inference LR test. With 7% mis-specification, the Wald statistic rejects 99% of the time (at the 95% confidence level) while the LR test rejects 22% of the time. At a sufficiently high degree of falseness both reject 100% of the time. Nonetheless, the LR test also has reasonable power. Figure 1, which provides histograms of the two test statistics for the true and a 3% false models, also shows the superior power of the Wald test. Figure 2, which shows the correlation coefficients between the two tests for the true and 3% false models, shows that there is little or no correlation between the two tests across samples. However, Figure 3, which is a scatter diagram of the correlations between the two test statistics on the same samples but for increasing degrees of falseness, shows that as the model becomes more false, both tests increase their rejection rate. Taken together, these findings suggest that, when one measure is well-fitting, it may be well-fitting or badly-fitting on the other measure. A possible explanation for these findings is that the two tests are measuring different things; the LR test is measuring the forecasting ability of the model while the Wald test is measuring the model's ability to explain the sample data.

Percent Mis-specified	Indirect Inference	Direct Inference
True	5.0	5.0
1	19.8	6.3
3	52.1	8.8
5	87.3	13.1
7	99.4	21.6
10	100.0	53.4
15	100.0	99.3
20	100.0	99.7

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

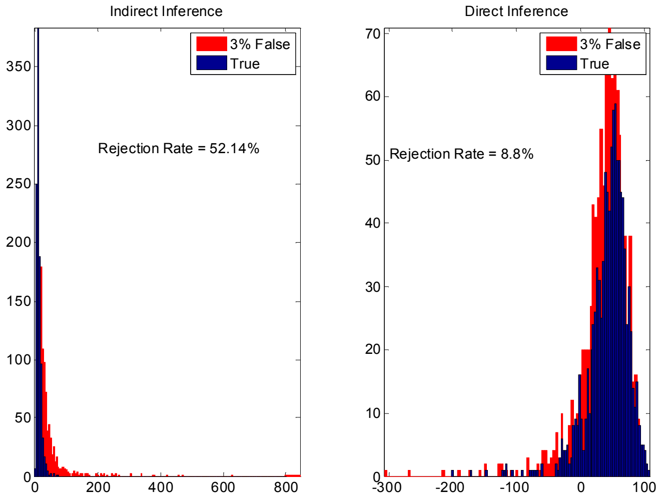

Figure 1: Histograms of true and 3% false models for Indirect and Direct Inference (some outliers deleted from Indirect for the chart)

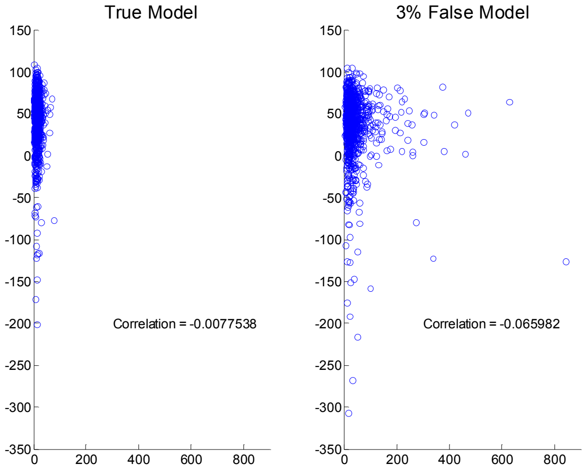

Figure 2: Scatter Plots of Indirect Inference (Wald) v. Direct Inference (LR) for 1000 samples of True Model (3 Variable VAR(1))

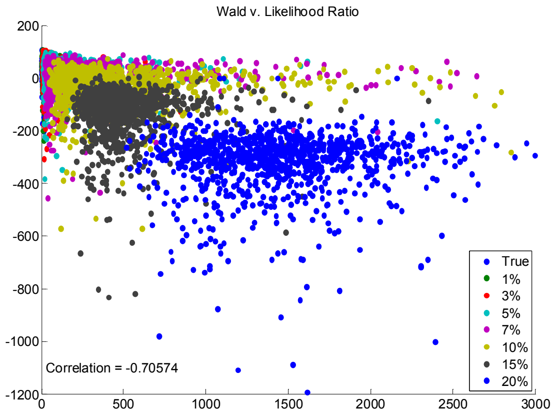

Figure 3: Scatter Plots of Indirect Inference (Wald) v. Direct Inference (LR) for True and False Models (some outliers taken out for clarity of scale)(3 Variable VAR(1))

4.1.1 Comparison of the tests with different VAR variable coverage and VAR lag order

Tests based on indirect inference that use VARs with a high-order of lags, or VARs with more than just a few variables, are extremely stringent and they tend to reject uniformly. In Le et al. (2011) we proposed 'directed' Wald tests where the information used in evaluating a DSGE model was deliberately reduced to cover only 'essential features' of the data; of course, all Wald tests are based on chosen features of the data and therefore are always to some degree 'directed'. Our use of the term is when the Wald test is focused on only a small subset of variables, or aspects of their behaviour.

We find in Table 2 that for the indirect inference test the power of the Wald statistic tends to rise as the number of variables in the VAR or its lag order is increased. But in Table 3 we find that the power of direct inference based on a Likelihood ratio test (using the LIML method on the residuals) does not appear to vary in any systematic way with the benchmark VAR used, either in terms of the number of variables included or the order of the VAR. The power functions are shown in Figure 4.

Why this is the case is a matter for future research. Our conjecture is that forecasting

VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1) — 9	5.00	19.76	52.14	87.30	99.38	100.00	100.00	100.00
3 variable VAR(2) — 18	5.00	38.24	68.56	84.10	99.64	100.00	100.00	100.00
3 variable VAR(3) — 27	5.00	38.22	65.56	92.28	99.30	100.00	100.00	100.00
5 variable VAR(1) — 25	5.00	28.40	77.54	97.18	99.78	100.00	100.00	100.00
7 variable VAR(3) — 147	5.00	75.10	99.16	99.96	100.00	100.00	100.00	100.00

Table 2: Indirect Inference Rejection Rates at 95% level for varying VARs

VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1) — 9	5.00	6.30	8.80	13.10	21.60	53.40	99.30	99.70
3 variable VAR(2) — 18	5.00	6.00	8.30	13.40	23.10	55.10	99.40	99.70
3 variable VAR(3) — 27	5.00	6.00	7.90	13.10	21.90	52.30	99.50	99.70
5 variable VAR(1) — 25	5.00	6.00	8.20	11.70	15.90	29.30	93.30	99.70
7 variable VAR(3) — 147	5.00	5.50	7.10	11.40	18.80	49.90	99.60	99.70

Table 3: Direct Inference Rejection Rates at 95% level for varying VARs

performance across different variables is highly correlated and that the most recent information provides the dominant input. If so, then adding variables or more lags would make little difference. With indirect inference the addition of variables or VAR detail adds to the complexity of behaviour that the DSGE model must match; the more complexity, the less well can the matching occur when the model is moderately false. Again, this brings out the essential difference in the two measures of performance.

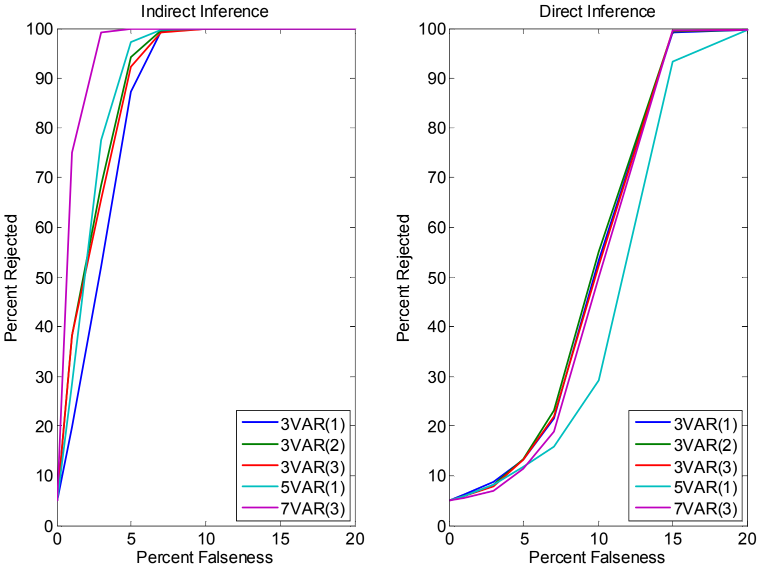

Figure 4: Power Functions for Indirect and Direct Inference for Various VARs.

4.1.2 Estimation and test power

In the above power comparisons we took the values of the DSGE model as given — perhaps by calibration or Bayesian estimation (where the priors may keep them away from the true values) or by some inefficient estimation process that fails to get close to the true parameter values. Suppose instead that we use maximum likelihood (FIML) estimates or indirect inference (II) estimates that minimise the Wald criterion. It is of interest to ask whether this would affect the previous power comparisons as we would then expect the model to be rejected only if it was mis-specified. For example, the model might assume Calvo price/wage setting when there was general competition or vice versa.

First, we examine the small sample properties of the two estimators. While we know from earlier work that the estimators have similar asymptotic properties, there is no work comparing their small sample properties. We assess the small sample bias of the two estimators using the same Monte Carlo experiment on the SW model. Thus, we endow the econometrician with the true general specification and re-estimate the model for each of the 1000 samples of data simulated from the true specification of the model. The percentage mean biases and the percentage absolute mean biases are reported in Table 4. We obtain a familiar result that the FIML estimates are heavily biased in small samples. By contrast, we find that the II estimator has very small bias; on average it is roughly half the FIML bias and the absolute mean bias is around 4%.

Second, we now check the power of each test for the re-estimated SW model against its general mis-specification which we require to be substantial otherwise the tests would have trivial power.[7] The type of mis-specification that we consider relates to the assumed degree of nominal rigidity in the model. The original SW model is New Keynesian (NK) with 100% Calvo contracting. An alternative specification is a New Classical (NC) version with 100%

[7] We can translate our results under re-estimation into terms of the 'degree of falseness' of the model as in the power functions used above. This will not be removed by the re-estimation process. Re-estimation will take the model's parameters to the corner solution where the estimates cannot get closer to the data without violating the model's general mis-specification.

			Mean Bias (%)		Absolute Mean Bias (%)	
		Starting coef	II	FIML	II	FIML
Steady-state elasticity of capital adjustment	φ	5.74	−0.900	5.297	0.900	5.297
Elasticity of consumption	σ_c	1.38	−5.804	−7.941	5.804	7.941
External habit formation	λ	0.71	−13.403	−21.240	13.403	21.240
Probability of not changing wages	ξ_w	0.70	−0.480	−3.671	0.480	3.671
Elasticity of labour supply	σ_L	1.83	0.759	−8.086	0.759	8.086
Probability of not changing prices	ξ_p	0.66	−1.776	0.027	1.776	0.027
Wage indexation	ι_w	0.58	−0.978	6.188	0.978	6.188
Price indexation	ι_p	0.24	0.483	3.228	0.483	3.228
Elasticity of capital utilisation	ψ	0.54	−13.056	−29.562	13.056	29.562
Share of fixed costs in production (+1)	Φ	1.50	−1.590	2.069	1.590	2.069
Taylor Rule response to inflation	r_p	2.04	7.820	2.815	7.820	2.815
Interest rate smoothing	ρ	0.81	−0.843	−0.089	0.843	0.089
Taylor Rule response to output	r_y	0.08	−4.686	−29.825	4.686	29.825
Taylor Rule response to change in output	$r_{\Delta y}$	0.22	−5.587	0.171	5.587	0.171
Average			−2.861	−5.758	4.155	8.586

Table 4: Small Sample Estimation Bias Comparison (II v. LR)

competitive markets and a one-quarter information lag about prices by households/workers. We then apply the II test of NC to data generated by NK, allowing full re-estimation by II for each sample and vice versa with a test of NK on data generated by NC. This is repeated using the LR test with re-estimation of each sample by FIML — technically we do this by minimising the LR on each sample.

	Percentage Rejected	
	NK data NC model	NC data NK model
II	99.6%	77.6%
LR	0%	0%

Table 5: Power of the test to reject a false model

The results in Table 5 strikingly confirm the relative lack of power of the LR test. On NK data, the rejection rate of the NC model with 95% confidence is 0%, and on NC data the rejection rate of the NK model is also 0%. It would seem, therefore, that with sufficient ingenuity the NC model can be re-estimated so as to forecast the data generated by the NK model even better than for the NK model itself (and vice versa) so that it is not rejected at all. By contrast when II is used, the power against general mis-specification is high. The NC model is rejected (with 95% confidence) 99.6% of the time on NK data and the NK model is rejected 78% of the time on NC data. The implication of this exercise is that the II test is indeed also far more powerful as a detector of general mis-specification than LR.

5 Extending the test comparison

We consider two extensions to the above experiments. First, instead of applying stationary shocks to the Smets-Wouters model as above, we apply non-stationary shocks. Second, partly in order to investigate whether these findings are model-specific, we carry out the same analysis, under both stationary and non-stationary shocks, to another widely-used DSGE model: the 3-equation (forward-looking IS curve, Phillips Curve and Taylor Rule) New Keynesian model of Clarida et al. (1999). We find that the previous conclusions do not change in any essential way for either model.

5.1 Non-stationary shocks applied to the SW model

If the data are non-stationary data then, in order to use the previous tests, we need to create an auxiliary model whose errors are stationary. We therefore use a VECM as the auxiliary model. Following Meenagh et al. (2012), and after log-linearisation, a DSGE model can usually be written in the form

$$A(L)y_t = BE_t y_{t+1} + C(L)x_t + D(L)e_t \tag{1}$$

where y_t are p endogenous variables and x_t are q exogenous variables which we assume are driven by

$$\Delta x_t = a(L)\Delta x_{t-1} + d + c(L)\epsilon_t. \tag{2}$$

The exogenous variables may consist of both observable and unobservable variables such as a technology shock. The disturbances e_t and ϵ_t are both iid variables with zero means. It follows that both y_t and x_t are non-stationary. L denotes the lag operator $z_{t-s} = L^s z_t$ and $A(L)$, $B(L)$ etc. are polynomial functions with roots outside the unit circle.

The general solution of y_t is

$$y_t = G(L)y_{t-1} + H(L)x_t + f + M(L)e_t + N(L)\epsilon_t. \tag{3}$$

where the polynomial functions have roots outside the unit circle. As y_t and x_t are non-stationary, the solution has the p cointegration relations

$$\begin{aligned} y_t &= [I - G(1)]^{-1}[H(1)x_t + f] \\ &= \Pi x_t + g. \end{aligned} \tag{4}$$

The long-run solution to the model is

$$\begin{aligned} \overline{y}_t &= \Pi \overline{x}_t + g \\ \overline{x}_t &= [1 - a(1)]^{-1}[dt + c(1)\xi_t] \\ \xi_t &= \Sigma_{i=0}^{t-1}\epsilon_{t-s}. \end{aligned}$$

Hence the long-run solution to x_t, namely, $\overline{x}_t = \overline{x}_t^D + \overline{x}_t^S$, has a deterministic trend $\overline{x}_t^D = [1 - a(1)]^{-1}dt$ and a stochastic trend $\overline{x}_t^S = [1 - a(1)]^{-1}c(1)\xi_t$.

The solution for y_t can therefore be re-written as the VECM

$$\begin{aligned} \Delta y_t &= -[I - G(1)](y_{t-1} - \Pi x_{t-1}) + P(L)\Delta y_{t-1} + Q(L)\Delta x_t + f + M(L)e_t + N(L)\epsilon_t \\ &= -[I - G(1)](y_{t-1} - \Pi x_{t-1}) + P(L)\Delta y_{t-1} + Q(L)\Delta x_t + f + \omega_t \qquad (5) \\ \omega_t &= M(L)e_t + N(L)\epsilon_t \end{aligned}$$

implying that, in general, the disturbance ω_t is a mixed moving average process. This suggests that the VECM can be approximated by the VARX

$$\Delta y_t = K(y_{t-1} - \Pi x_{t-1}) + R(L)\Delta y_{t-1} + S(L)\Delta x_t + g + \zeta_t \tag{6}$$

where ζ_t is an iid zero-mean process. As

$$\overline{x}_t = \overline{x}_{t-1} + [1 - a(1)]^{-1}[d + \epsilon_t]$$

the VECM can also be written as

$$\Delta y_t = K[(y_{t-1} - \overline{y}_{t-1}) - \Pi(x_{t-1} - \overline{x}_{t-1})] + R(L)\Delta y_{t-1} + S(L)\Delta x_t + h + \zeta_t. \tag{7}$$

Either of equations (6) or (7) can act as the auxiliary model. Here we focus on equation (7) which distinguishes between the effect of the trend component of x_t and the temporary deviation of x_t from trend. These two components have different effects in our models and so should be distinguished in the data in order to allow the tests to provide the fullest discrimination. It is possible to estimate equation (7) in one stage by OLS. Using Monte Carlo experiments, Meenagh et al. (2012) show that this procedure is extremely accurate. We therefore use this auxiliary model as our benchmark both for the II test and the LR test.

To generate non-stationary data from the DSGE model we endow it with one or more non-stationary error processes. These are constructed by generating AR processes for differences in the structural errors. For the SW model we add banking and money and give it a non-stationary productivity shock. Full details of this version of the SW model are in Le, Meenagh and Minford (2012). The rejection probabilities for the Wald and LR tests are reported respectively in Tables 6 and 7. Once more the test based on indirect inference has far more power than the direct LR test.

VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1) — 9	5.0	7.9	49.2	97.8	100.0	100.0	100.0	100.0
3 variable VAR(2) — 18	5.0	9.2	45.0	99.2	100.0	100.0	100.0	100.0
3 variable VAR(3) — 27	5.0	7.1	40.5	98.6	100.0	100.0	100.0	100.0
5 variable VAR(1) — 25	5.0	11.1	57.9	99.6	100.0	100.0	100.0	100.0
7 variable VAR(3) — 147	5.0	19.9	77.4	100.0	100.0	100.0	100.0	100.0

Table 6: Indirect Inference Rejection Rates at 95% level for varying VARs (non-stationary data)

VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1) — 9	5.0	5.2	5.8	6.2	7.4	9.6	15.6	26.5
3 variable VAR(2) — 18	5.0	5.1	5.8	6.0	7.3	9.4	15.1	26.2
3 variable VAR(3) — 27	5.0	5.3	5.8	6.1	7.3	9.5	15.5	26.3
5 variable VAR(1) — 25	5.0	5.7	6.1	7.2	7.9	9.6	12.6	21.6
7 variable VAR(3) — 147	5.0	5.0	6.0	7.1	8.3	10.7	15.0	25.3

Table 7: Direct Inference Rejection Rates at 95% level for varying VARs (non-stationary data)

5.2 Extension to the 3-equation New Keynesian model

The results for the 3-equation New Keynesian model are reported for stationary data in Tables 8 and for non-stationary data in Tables 9. The results are not much different from those for the much larger Smets-Wouters model. For stationary data the power of the indirect inference test rises rapidly with the degree of falseness, but that of the Likelihood Ratio is much poorer and rises less fast. For non-stationary data the power of the indirect inference test rises less fast than for the Smets-Wouters model, while the power of the LR test is very low and hardly increases with the degree of falseness.

These findings suggest that, if one is only interested in these three major macro variables, there is no substantial power penalty in moving to a more aggregative model of the economy if indirect inference is used. The power of the LR test is also similar for the two models — but lower than the Wald test — for stationary data and much lower for non-stationary data.

INDIRECT INFERENCE								
VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1) — 4	5.0	16.8	82.6	99.6	100.0	100.0	100.0	100.0
3 variable VAR(1) — 9	5.0	25.1	97.7	100.0	100.0	100.0	100.0	100.0
3 variable VAR(2) — 18	5.0	16.1	77.2	98.4	100.0	100.0	100.0	100.0
3 variable VAR(3) — 27	5.0	14.4	73.0	97.5	99.7	100.0	100.0	100.0

DIRECT INFERENCE								
VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1) — 4	5.0	6.0	7.5	9.9	13.2	18.7	29.2	39.3
3 variable VAR(1) — 9	5.0	5.2	6.9	9.0	12.3	18.8	32.3	51.3
3 variable VAR(2) — 18	5.0	5.7	7.2	10.3	13.0	18.8	32.8	51.6
3 variable VAR(3) — 27	5.0	5.4	7.4	9.6	12.3	19.1	33.0	51.6

Table 8: 3-EQUATION MODEL: STATIONARY data: Rejection Rates at 95% level for varying VARs

INDIRECT INFERENCE								
VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1) — 4	5.0	9.6	35.6	78.6	93.6	100.0	100.0	100.0
3 variable VAR(1) — 9	5.0	2.9	9.4	40.6	63.1	99.4	100.0	100.0
3 variable VAR(2) — 18	5.0	3.7	12.0	34.8	62.8	96.8	100.0	100.0
3 variable VAR(3) — 27	5.0	3.1	10.8	34.7	55.3	96.9	100.0	100.0

DIRECT INFERENCE								
VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1) — 4	5.0	5.3	5.4	5.6	6.3	7.5	9.2	10.7
3 variable VAR(1) — 9	5.0	5.2	5.3	5.5	5.5	5.7	5.7	5.9
3 variable VAR(2) — 18	5.0	5.2	5.3	5.5	5.5	5.7	5.7	5.9
3 variable VAR(3) — 27	5.0	5.2	5.3	5.5	5.5	5.7	5.7	5.9

Table 9: 3-EQUATION MODEL: NON-STATIONARY DATA: Indirect Inference Rejection Rates at 95% level for varying VARs

6 Why does the indirect inference test have greater power than the Likelihood Ratio test?

We have found that a likelihood-based test of a DSGE model may deliver weak power compared with a test based on indirect inference. This could be due to a flat likelihood surface. Canova and Sala (2009) suggested that this may occur when the model is poorly identified. Strictly speaking, identification is a theoretical property of a (DSGE) model; it occurs when the reduced form can also be generated by a different model. However, using indirect inference, Le, Minford and Wickens (2013) tested both of the DSGE models analysed here to see whether these models generated data whose reduced form (or approximations to it) could also be generated by other DSGE models. They concluded that both of the models are highly over-identified. Had it been possible to find such an alternative model, the percentage of rejections would been the same as for the true original model. The nearest alternative models they were able to find were rejected nearly 100% of the time for a 5% significance test when, of course, the true model is only rejected 5% of the time.

Modelling the structural disturbances of DSGE models as autogressive processes is clearly likely to improve their overall fit to the data. It may also be a cause of the flatness of the likelihood surface of DSGE models. Taken together with the fact that when we create a

False model we also change the original autocorrelation structure of the disturbances and so for the LR test we estimate new autogressive processes, we conjecture that this re-estimation may disguise false structural parameterisations and so contribute to LR tests having weak power. This is because a false parameterisation will generate false structural error processes and the autoregressive processes fitted to these false errors will tend to bring the model back on track and, in particular, to improve the model's forecasts of next-period outcomes, which is what the likelihood function is based on.

To check whether re-estimating the errors affects the power properties of the LR test we calculate its power using the AR coefficients of the original structural errors and those from altering the AR coefficients to fit the errors of the false model. The difference in the powers represents the additional power due to re-estimating the AR coefficients. We carry out the calculations for both of the DSGE models discussed above. The results are reported in Tables 10 and 11. We find that the power of the LR test is similar for the two error processes for both DSGE models; in some cases the power is even reduced as a result of re-estimating the AR coefficients. The finding that re-estimating the AR coefficients has little effect on the power of the LR test is consistent with the likelihood surface being flat and little affected by different values of the AR coefficients, as is the weak increase in power when the structural coefficients are falsified.

In contrast to the LR test, indirect inference does not use tracking performance as a measure of 'fit'. Instead, it compares the reduced form as captured by the auxiliary model from data simulated from the model with the actual historical data. In effect, the test is exploiting the DSGE model's over-identification which causes the True model and the false models to have quite different reduced forms. These reduced forms reflect not only the structural parameters of the model but also all the properties of the structural error processes, their AR coefficients and the skewness, kurtosis and covariances of the innovations in the structural errors processes. Differences in any of these features between the reduced forms for the True model and the false models may potentially cause rejection.

RE-ESTIMATED AR COEFFS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1)	6.3	8.8	13.1	21.6	53.4	99.3	100.0
3 variable VAR(2)	6.0	8.3	13.4	23.1	55.5	99.4	100.0
3 variable VAR(3)	6.0	7.9	13.1	21.9	52.3	99.5	100.0
5 variable VAR(1)	6.0	8.2	11.7	15.9	29.3	93.3	100.0
7 variable VAR(3)	5.5	7.1	11.4	18.8	49.9	99.6	100.0
ORIGINAL AR COEFFS							
3 variable VAR(1)	5.3	6.5	9.7	18.9	52.0	98.5	100.0
3 variable VAR(2)	4.4	5.4	9.0	18.2	52.0	98.3	100.0
3 variable VAR(3)	4.8	6.0	9.9	19.7	52.5	98.5	100.0
5 variable VAR(1)	5.2	6.6	8.7	14.2	40.4	98.0	100.0
7 variable VAR(3)	4.4	6.5	13.2	30.9	78.3	99.7	100.0
DIFFERENCE							
3 variable VAR(1)	1.0	2.3	3.4	2.7	1.4	0.8	0.0
3 variable VAR(2)	1.6	2.9	4.4	4.9	3.1	1.1	0.0
3 variable VAR(3)	1.2	1.9	3.2	2.2	−0.2	1.0	0.0
5 variable VAR(1)	0.8	1.6	3.0	1.7	−11.1	−4.7	0.0
7 variable VAR(3)	1.1	0.6	−1.8	−12.1	−28.4	−0.1	0.0

Table 10: SW MODEL, STATIONARY DATA: Decomposition of the power of LR

RE-ESTIMATED AR COEFFS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1)	5.0	5.7	6.5	7.7	9.4	11.9	13.5
3 variable VAR(1)	4.9	5.0	5.6	6.1	7.9	10.9	13.5
3 variable VAR(2)	5.0	5.1	5.6	6.2	7.7	11.2	13.8
3 variable VAR(3)	5.3	5.1	5.9	6.8	8.0	11.5	14.3
ORIGINAL AR COEFFS							
2 variable VAR(1)	4.8	4.9	5.5	5.3	7.6	11.5	16.5
3 variable VAR(1)	4.4	4.3	3.8	4.0	5.3	8.4	14.7
3 variable VAR(2)	4.6	4.2	4.2	4.8	6.1	8.9	15.3
3 variable VAR(3)	4.7	4.2	4.2	4.8	6.3	9.8	15.3
DIFFERENCE							
2 variable VAR(1)	0.2	0.8	1.0	2.4	1.8	0.4	−3.0
3 variable VAR(1)	0.5	0.7	1.8	2.1	2.6	2.5	−1.2
3 variable VAR(2)	0.4	0.9	1.4	1.4	1.6	2.3	−1.5
3 variable VAR(3)	0.6	0.9	1.7	2.0	1.7	1.7	−1.0

Table 11: 3-EQUATION MODEL, STATIONARY DATA: Decomposition of the power of LR

The contributions to the power of the Wald test due to these different types of falseness can be checked. In particular, we examine the effects on power of falseness of three elements: the structural parameters, the AR coefficients of the structural error processes and the moment properties of the innovations in the structural errors. The results reported in Tables 12 and 13 are the powers when all three elements are falsified and when each element on its own is falsified.

ALL ELEMENTS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1)	19.8	52.1	87.3	99.4	100.0	100.0	100.0
3 variable VAR(2)	38.2	68.6	84.1	99.6	100.0	100.0	100.0
3 variable VAR(3)	38.2	65.6	92.3	99.3	100.0	100.0	100.0
5 variable VAR(1)	28.4	77.5	97.2	99.8	100.0	100.0	100.0
7 variable VAR(3)	75.1	99.2	100.0	100.0	100.0	100.0	100.0
STRUCTURAL PARAMETERS							
3 variable VAR(1)	7.4	14.0	40.1	68.4	98.2	100.0	100.0
3 variable VAR(2)	7.2	10.4	32.1	56.1	92.5	100.0	100.0
3 variable VAR(3)	7.8	12.8	30.8	51.0	88.8	100.0	100.0
5 variable VAR(1)	7.3	12.8	26.6	58.0	96.3	100.0	100.0
7 variable VAR(3)	53.1	93.0	99.5	100.0	100.0	100.0	100.0
AR PARAMETERS							
3 variable VAR(1)	10.4	33.6	64.1	88.3	98.8	100.0	100.0
3 variable VAR(2)	10.0	31.9	63.4	87.9	98.4	99.9	100.0
3 variable VAR(3)	11.8	32.9	62.6	86.4	98.5	99.9	100.0
5 variable VAR(1)	19.6	80.1	98.3	99.9	100.0	100.0	100.0
7 variable VAR(3)	65.0	99.2	100.0	100.0	100.0	100.0	100.0
SHOCKS							
3 variable VAR(1)	5.4	4.7	4.3	5.5	5.1	5.5	4.8
3 variable VAR(2)	6.3	4.5	5.6	4.9	5.5	4.5	4.6
3 variable VAR(3)	7.6	6.1	5.5	6.3	6.7	6.5	5.6
5 variable VAR(1)	6.2	6.1	4.5	6.2	6.8	6.0	5.2
7 variable VAR(3)	31.6	33.9	26.4	27.9	28.0	25.3	28.1

Table 12: SW MODEL, STATIONARY DATA: Decomposition of the power of II

For both models the power of the Wald test arises mainly from false structural parameters and autocorrelation coefficients. Each on its own has high power and taken together they generate very high power. But the power arising from the innovation processes (the shocks) is much less, especially for the SW model. The finding of high power for the indirect inference procedure has also been found to be robust to the choice of auxiliary model; neither varying the number of variables nor the order of the lags in the VAR greatly affects its power. For example, the Wald test has high power for the SW model even using a VAR(1) with 3

ALL ELEMENTS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1)	16.8	82.6	99.6	100.0	100.0	100.0	100.0
3 variable VAR(1)	25.1	97.7	100.0	100.0	100.0	100.0	100.0
3 variable VAR(2)	16.1	77.2	98.4	100.0	100.0	100.0	100.0
3 variable VAR(3)	14.4	73.0	97.5	99.7	100.0	100.0	100.0
STRUCTURAL PARAMETERS							
2 variable VAR(1)	7.3	8.7	12.6	19.3	40.4	76.1	92.7
3 variable VAR(1)	6.2	10.1	25.5	53.7	80.7	99.4	100.0
3 variable VAR(2)	6.8	9.3	12.8	20.6	45.9	77.2	95.0
3 variable VAR(3)	5.8	7.5	12.0	21.7	45.8	74.0	95.5
AR PARAMETERS							
2 variable VAR(1)	16.2	86.3	99.7	100.0	100.0	100.0	100.0
3 variable VAR(1)	18.8	96.8	100.0	100.0	100.0	100.0	100.0
3 variable VAR(2)	16.5	87.3	99.9	100.0	100.0	100.0	100.0
3 variable VAR(3)	18.9	81.6	99.5	100.0	100.0	100.0	100.0
SHOCKS							
2 variable VAR(1)	5.6	6.8	5.7	10.1	15.0	27.3	46.7
3 variable VAR(1)	5.4	6.0	8.4	8.7	11.7	26.7	48.8
3 variable VAR(2)	5.6	5.4	5.1	9.0	13.1	31.0	41.8
3 variable VAR(3)	4.9	6.1	4.1	9.0	12.4	29.5	48.2

Table 13: 3-EQUATION MODEL, STATIONARY DATA: Decomposition of the power of II

variables as the auxiliary model when the true reduced form of the SW model is likely to be a VAR of 7 variables with an order 2 or higher. These different VAR representations each reflect sufficient constituent elements of these DSGE models and their structural errors to enable them to distinguish starkly between True and false parameterisations of these DSGE models.

6.1 Non-stationary data power decomposition

This power decomposition can be repeated with non-stationary data. The results for the LR test for the two models are reported in Tables 14 and 15. The power of the LR test is much lower when the data are non-stationary, while re-estimating the AR coefficients causes a further small reduction in power. A possible explanation is that with non-stationary data the estimation error introduced by false AR coefficients matters less for forecasting performance. This may be because the power of the test will be determined more by the long-run structural parameters than by the short-run dynamic behaviour of the structural errors.

RE-ESTIMATED AR COEFFS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1)	5.1	4.6	4.1	5.8	13.4	30.9	45.0
3 variable VAR(2)	4.9	4.4	4.0	5.6	12.4	28.9	43.9
3 variable VAR(3)	4.8	4.3	3.9	5.4	12.4	28.7	43.5
5 variable VAR(1)	5.7	7.8	10.5	12.6	17.3	27.1	37.3
7 variable VAR(3)	5.5	6.7	8.2	10.4	14.8	26.1	38.8
ORIGINAL AR COEFFS							
3 variable VAR(1)	5.4	5.2	5.4	8.2	16.8	34.2	49.0
3 variable VAR(2)	5.4	5.3	5.4	8.2	27.0	34.4	48.5
3 variable VAR(3)	5.5	5.3	5.7	8.1	17.4	34.5	49.5
5 variable VAR(1)	5.6	7.6	9.9	14.4	22.8	38.5	51.1
7 variable VAR(3)	6.0	7.0	9.2	11.6	17.8	34.2	52.4
DIFFERENCE							
3 variable VAR(1)	−0.3	−0.6	−1.3	−2.4	−3.4	−3.3	−4.0
3 variable VAR(2)	−0.5	−0.9	−1.4	−2.5	−4.6	−5.5	−4.6
3 variable VAR(3)	−0.7	−1.0	−1.8	−2.7	−5.0	−5.8	−6.0
5 variable VAR(1)	0.1	0.2	0.6	−1.8	−5.5	−11.4	−13.8
7 variable VAR(3)	−0.5	−0.3	−1.0	−1.2	−3.0	−8.1	−13.6

Table 14: SW MODEL, NON-STATIONARY DATA: Decomposition of the power of LR

RE-ESTIMATED RHOS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1)	5.3	5.4	5.6	6.3	7.5	9.2	10.7
3 variable VAR(1)	5.2	5.3	5.5	5.5	5.7	5.7	5.9
3 variable VAR(2)	5.2	5.3	5.5	5.5	5.7	5.7	5.9
3 variable VAR(3)	5.2	5.3	5.5	5.5	5.7	5.7	5.9
TRUE RHOS							
2 variable VAR(1)	5.2	6.1	6.9	7.4	8.4	10.6	12.8
3 variable VAR(1)	5.3	5.7	6.0	6.0	6.1	6.2	6.5
3 variable VAR(2)	5.3	5.7	6.0	6.0	6.1	6.2	6.5
3 variable VAR(3)	5.3	5.7	6.0	6.0	6.1	6.2	6.5
DIFFERENCE							
2 variable VAR(1)	0.1	−0.7	−1.3	−1.1	−0.9	−1.4	−2.1
3 variable VAR(1)	−0.1	−0.4	−0.5	−0.5	−0.4	−0.5	−0.6
3 variable VAR(2)	−0.1	−0.4	−0.5	−0.5	−0.4	−0.5	−0.6
3 variable VAR(3)	−0.1	−0.4	−0.5	−0.5	−0.4	−0.5	−0.6

Table 15: 3-EQUATION MODEL, NON-STATIONARY DATA: Decomposition of the power of LR

The result for the Indirect Inference test are reported in Tables 16 and 17. We find that the role of the error innovations increases in importance when the data are non-stationary compared with when they are stationary. This may be because the effect of wrong innovations in the non-stationary errors is permanent, thus worsening the model's dynamics and volatility compared with those of the true data processes.

Overall, the contrast between the two methods is enhanced when the data are non-stationary. Under LR the AR coefficients attenuate the power of the test; while under II the power of the test is increased both by the AR coefficients and by the innovations' properties. This highlights the importance in the II test of the properties of the error processes in the model. A DSGE model consists of two main parts: its structural coefficients and its shock processes. Both are tested in Indirect Inference.

ALL ELEMENTS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1)	4.9	39.9	98.6	100.0	100.0	100.0	100.0
3 variable VAR(2)	6.6	45.5	99.4	100.0	100.0	100.0	100.0
3 variable VAR(3)	5.6	42.6	99.3	100.0	100.0	100.0	100.0
5 variable VAR(1)	7.9	49.7	99.5	100.0	100.0	100.0	100.0
7 variable VAR(3)	10.6	69.1	100.0	100.0	100.0	100.0	100.0
STRUCTURAL PARAMETERS							
3 variable VAR(1)	4.6	23.6	82.6	99.9	100.0	100.0	100.0
3 variable VAR(2)	5.1	30.3	90.5	99.9	100.0	100.0	100.0
3 variable VAR(3)	5.3	28.4	85.9	99.9	100.0	100.0	100.0
5 variable VAR(1)	5.9	23.6	79.8	99.5	100.0	100.0	100.0
7 variable VAR(3)	8.8	43.7	97.6	100.0	100.0	100.0	100.0
AR PARAMETERS							
3 variable VAR(1)	4.6	7.2	11.3	15.9	25.4	50.0	71.2
3 variable VAR(2)	4.4	5.4	8.1	12.8	22.3	49.6	73.5
3 variable VAR(3)	4.3	5.0	6.0	10.0	17.0	34.7	57.6
5 variable VAR(1)	6.4	16.0	37.7	58.0	66.9	80.5	90.4
7 variable VAR(3)	8.5	19.3	48.0	74.7	84.9	94.7	98.9
SHOCKS							
3 variable VAR(1)	4.7	4.2	5.0	6.6	9.9	20.8	47.4
3 variable VAR(2)	4.6	5.0	5.6	8.0	10.7	27.9	58.4
3 variable VAR(3)	4.8	4.6	5.9	7.7	10.8	17.0	55.6
5 variable VAR(1)	5.4	5.0	5.6	6.0	6.5	11.6	18.7
7 variable VAR(3)	6.5	6.9	7.1	7.2	8.2	12.8	18.1

Table 16: SW MODEL, NON-STATIONARY DATA: Decomposition of the power of II

ALL ELEMENTS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1)	8.1	40.8	81.4	96.3	100.0	100.0	100.0
3 variable VAR(1)	4.7	11.7	31.3	69.1	99.1	100.0	100.0
3 variable VAR(2)	2.3	11.5	35.0	51.1	95.1	100.0	100.0
3 variable VAR(3)	3.2	9.4	33.7	49.1	95.0	100.0	100.0
STRUCTURAL PARAMETERS							
2 variable VAR(1)	6.1	7.2	9.4	15.2	28.6	54.7	80.6
3 variable VAR(1)	6.9	7.8	10.9	14.9	27.0	46.0	77.7
3 variable VAR(2)	6.4	9.0	12.8	16.6	31.9	54.7	85.8
3 variable VAR(3)	6.7	8.3	11.6	16.0	27.2	54.0	85.6
AR PARAMETERS							
2 variable VAR(1)	9.3	50.6	90.2	99.2	100.0	100.0	100.0
3 variable VAR(1)	3.9	15.5	64.5	90.6	99.9	100.0	100.0
3 variable VAR(2)	3.8	26.3	67.6	91.6	99.8	100.0	100.0
3 variable VAR(3)	3.4	26.3	70.6	90.9	100.0	100.0	100.0
SHOCKS							
2 variable VAR(1)	5.9	5.1	6.9	8.0	10.9	30.4	47.4
3 variable VAR(1)	5.4	6.0	7.9	6.6	8.2	13.8	30.2
3 variable VAR(2)	5.6	5.6	7.4	6.7	9.0	23.1	32.8
3 variable VAR(3)	5.4	6.0	6.5	6.7	8.3	24.6	27.0

Table 17: 3-EQUATION MODEL, NON-STATIONARY DATA: Decomposition of the power of II

6.2 Comparing tests of the SW model post-1984 as re-estimated by Le et al. (2011) and others

Le et al. (2011) found that, after re-estimation by indirect inference, the SW model on post-1984 (but pre-crisis) data passed the indirect inference test comfortably. It is of interest to examine the outcome from using a likelihood ratio test. The II test used a VAR(1) with three variables — output, inflation and interest rates — as the auxiliary model. With a higher-order VAR for these 3 variables, as well as with a VAR(1) with more than these three variables, the model performed progressively worse, being rejected most of the time. Le et al. interpreted this to mean that the model is able to capture the 'broad outlines' of the behaviour of these key macroeconomic variables but the model is not the full 'truth'.

We choose as the benchmark for the LR test a VAR(1) with 3 variables, as we have seen that the power does not vary with the lag order of the VAR or with the number of variables. For both the LR and Wald tests we generate 1000 sets of bootstrap data from the model's errors from which we obtain critical values from estimates of the distributions of the test statistics under the null that the model is true. The probabilities of rejecting the null that the model is correct and the VAR is restricted against the alternative of an unrestricted VAR are reported in Table 18.

We have found in our Monte Carlo experiments that the power of the LR test is considerably lower than that for the Wald test; with more variables in the VAR and with higher-order lags, we found that the power of the Wald test rose substantially, while remaining little changed for the direct inference LR test. This is consistent with what we find here for the modified SW model. Of the two tests, the LR fails to reject at all, while the Wald rejects for any VAR with more than 18 coefficients. We can also see that for our main focus on three variables with a VAR(1) (the first line of Table 20) both tests give consistent results.

VAR — no. of coefficients	Wald^{+}	LR
3 variable VAR(1) — 9	83.5	71.7
3 variable VAR(2) — 18	99.6	71.4
3 variable VAR(3) — 27	100	67.7
4 variable VAR(1) — 16	90.1	82.8
5 variable VAR(1) — 25	96.6	74.2
7 variable VAR(3) — 147	100	13.4

$^{+}$The Wald test includes the variances of the data in each case

Table 18: Tests using varying VARs

Comparing the outcomes for the two tests, the LR tests are all passed rather easily indicating that the model is well 'on track'. This was noticed by Smets and Wouters for their original model on which they performed various forecasting tests that are closely related to the LR test used here. In contrast, the model passes the Wald test only using a VAR(1) with 3 or 4 key variables, which is a coarse description of the inter-relationships. For finer descriptions or with more variables, the model fails. This provides information about what the model can do. In general we find that macro models cannot match the details of consumption and investment, even when they can match the key variables: output, inflation and interest rates. A possible reason is that the data on consumption and investment are poor; for example, we know that durable consumption goods, which should be treated as capital, are routinely included in consumption.

Table 19 summarises the results of many of the recent applications of the use of the indirect inference evaluation procedure. The Wald statistic used is based on the coefficients of the auxilary VAR model and the data variances. The first column denotes the country, sample episode and the model studied; the third column provides the name of the authors and the reference. The second column gives the results which show that models can be found that are not rejected for key sets of macro variables such as output, inflation and interest rates. The findings of Le et al (2010, 2011, 2014) are that, in general, models which can match a VAR(1) on a limited number of variables, do not perform as well on VARs with many more variables, and are typically rejected for higher-order VARs than a VAR(1).

Another common finding is that the 3-equation New Keynesian model originally proposed by Clarida, Gali and Gertler (1999) passes the test after re-estimation and can even match higher-order VARs - see Minford and Ou (2013), Liu and Minford (2012, 2014), and also Minford, Ou and Wickens (2013) for similar results. A possible explanation is the relative lack of tight cross-equation restrictions in these small models compared with those imposed by the more elaborate model of Smets and Wouters.

Country	Episode	Model	Estimation method	Result/Wald %tile[†]	references
UK	1975–2004	Liverpool Model (3 regimes)	Calibrated	Marginal/98.8	Minford et al. (2009)
EA	1975–1999	Smets-Wouters	Bayesian	Reject/100	Meenagh et al. (2009)
EA+US	1975–1999	Smets-Wouters ‘world’	Bayesian	Outputs,RXR/94.2	Le et al. (2010)
US	1982–2007	3-eqn NK-M[⊙]	Calibrated	$y, \pi, R/96.5$	Minford and Ou (2013)
UK	1959–2007	RBC open economy[*]	Calibrated	$RXR/94.2$	Meenagh et al. (2010)
US	1947–2004	Smets-Wouters hybrid[⊺]	Indirect estimation	$y, \pi, R/98.7$	Le et al. (2011)
US	1984–2004	Smets-Wouters hybrid[⊺]	Indirect estimation	$y, \pi, R/83.8$	Le et al. (2011)
US	1981–2010	3-eqn NK (Rational Exp.)	Indirect estimation	$y, \pi, R/79.8$	Liu and Minford (2014a)
US	1981–2010	3-eqn NK (Behavioural Exp.)	Indirect estimation	Reject/100	Liu and Minford (2014a)
US	1981–2010	4-eqn NK+banking[‡]	Indirect estimation	$y, \pi, R/45.4$	Liu and Minford (2014b)
China	1978–2007	Smets-Wouters hybrid[⊺]	Indirect estimation	$y, \pi, R/69.0$	Dai et al. (2014)
China	1991–2011	Smets-Wouters hybrid+bkg[⊺][‡][*]	Indirect estimation	$y, \pi, R/89.2$	Le et al. (2014)

[*] non-stationary data

[‡] Addition of the banking sector model of Bernanke et al. (1999)

[⊙] New Keynesian with imposition of timeless monetary policy rule

[⊺] Smets-Wouters with addition of competitive sector

[†] Results column shows variables included in Wald and Wald rejection status with Wald percentile.

Table 19: Summary of recent tests of DSGE models.

7 Conclusions

In comparing the Indirect Inference test with the Likelihood Ratio test we have noted that the two tests measure the performance of a model in different ways: the LR test, a direct test, measures its forecasting performance, and the Wald test, which is an example of indirect inference, measures a model's ability to replicate descriptors of the data. While both tests have increased power in small samples as the degree of falseness of a model increases, their outcomes may be dissimilar across different models. It follows that the Wald and LR tests are measuring quite different things.

The main conclusion of this paper based on two DSGE models, the Smets-Wouters model and a three-equation New Keynesian model, is that Indirect Inference is a powerful test of model specification while the direct inference Likelihood Ratio test has much less power. The reason for this difference appears to lie in the treatment of model structural errors. In the LR test the model-implied error processes can, and invariably are, re-estimated so that the model's forecasting performance can be partially preserved by this re-estimation; the structural innovations are then discarded in the test. In contrast, in the II test the model-implied structural errors are a key element of the Wald test. Both tests use the coefficients of the time series processes generating the structural errors but the II test uses, in addition, the matrix of moments of the innovations to the structural error processes. These contribute, sometimes in an important way, to the power of the II test, whereas the power of the LR test rests almost entirely on the structural coefficients. Presumably, for similar reasons, the Indirect estimator suffers much less small-sample bias than the ML estimator.

Our results, especially those based on Indirect Inference, suggest that researchers need to focus carefully on the features of the data they wish their models to explain otherwise the model is likely to be rejected. This may be why early DSGE models were being rejected by LR tests; probably, these models did not have sufficient dynamic content. The more recent SW model does pass these tests for the post-1984 sample when re-estimated by indirect methods.

In summary, the properties of tests of DSGE models based on indirect inference and small samples encourage its further use.

References

[1] Basawa, I. V., Mallik, A. K., McCormick, W. P., Reeves, J. H., and Taylor, R. L. (1991), "Bootstrapping unstable first-order autoregressive processes," *Annals of Statistics,* 19, 1098–1101.

[2] Bernanke, B. S., Gertler, M., and Gilchrist, S. (1999), "The financial accelerator in a business cycle framework," *Handbook of Macroeconomics*, vol. 1, edited by J.B. Taylor and M. Woodford, ch. 21, 1341–1393, Elsevier.

[3] Canova, F. (1994), "Statistical Inference in Calibrated Models," *Journal of Applied Econometrics,* 9, S123–144.

[4] Canova, F. (1995), "Sensitivity Analysis and Model Evaluation in Dynamic Stochastic General Equilibrium Models," *International Economic Review,* 36, 477–501.

[5] Canova, F. (2005), "Methods for Applied Macroeconomic Research," Princeton University Press, Princeton.

[6] Canova, F., and Sala, L. (2009), "Back to square one: Identification issues in DSGE models," *Journal of Monetary Economics,* 56, 431–449.

[7] Christiano, L. (2007), Comment on "On the fit of new Keynesian models" by Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. *Journal of Business and Economic Statistics,* 25,143–151.

[8] Christiano, L. J., Eichenbaum, M., and Evans, C. L. (2005), "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113(1), 1-45.

[9] Clarida, R., Gali, J. and Gertler, M. L. (1999), "The Science of Monetary Policy: A New Keynesian Perspective," *Journal of Economic Literature*, 37(4), pp.1661-1707.

[10] Dai, L., Minford, P. and Zhou, P. (2014), "A DSGE model of China," *Cardiff Economics Working Paper* No E2014/4, Cardiff University, Cardiff Business School, Economics Section; also CEPR discussion paper 10238, CEPR, London.

[11] Dave,C., and De Jong, D. N. (2007), "Structural Macroeconomics," Princeton University Press.

[12] Del Negro, M., and Schorfheide, F. (2004), "Priors from General Equilibrium Models for VARs," *International Economic Review*, 45, 643–673.

[13] Del Negro, M., and Schorfheide, F. (2006), "How good is what you've got? DSGE-VAR as a toolkit for evaluating DSGE models," *Economic Review*, Federal Reserve Bank of Atlanta, issue Q2, 21–37.

[14] Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. (2007a), "On the fit of new Keynesian models," *Journal of Business and Economic Statistics,* 25,123–143.

[15] Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. (2007b), Rejoinder to Comments on "On the fit of new Keynesian models," by Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. *Journal of Business and Economic Statistics*, 25,159–162.

[16] Evans, R. and Honkapohja, S. (2005), "Interview with Thomas J. Sargent," *Macroeconomic Dynamics*, 9, 2005, 561–583.

[17] Faust, J. (2007), Comment on "On the fit of new Keynesian models," by Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R., *Journal of Business and Economic Statistics,* 25,154–156.

[18] Fernandez-Villaverde, J., Rubio-Ramirez, F., Sargent, T, and Watson, M. (2007), "ABCs (and Ds) of Understanding VARs," *American Economic Review*, pp 1021-1026.

[19] Friedman, M. (1953), "The methodology of positive economics," *Essays in Positive Economics*, Chicago: University of Chicago Press.

[20] Gallant, A. R. (2007), Comment on "On the fit of new Keynesian models," by Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R., *Journal of Business and Economic Statistics,* 25,151–152.

[21] Gourieroux, C., and Monfort, A. (1995), "Simulation Based Econometric Methods," *CORE Lectures Series*, Louvain-la-Neuve.

[22] Gourieroux, C., Monfort, A., and Renault, E. (1993), "Indirect inference," *Journal of Applied Econometrics,* 8, S85–S118.

[23] Gregory, A., and Smith, G. (1991), "Calibration as testing: Inference in simulated macro models," *Journal of Business and Economic Statistics,* 9, 293–303.

[24] Gregory, A., and Smith, G. (1993), "Calibration in macroeconomics," in: Maddala, G. (Ed.), *Handbook of Statistics* vol. 11, Elsevier, St. Louis, Mo., pp. 703–719.

[25] Hansen, B. E. (1999), "The Grid Bootstrap And The Autoregressive Model," *The Review of Economics and Statistics,* 81, 594–607.

[26] Hansen, L. P. and Heckman, J. J. (1996), "The empirical foundations of calibration," *Journal of Economic Perspectives,* 10(1):87–104.

[27] Horowitz, J. L. (2001a), "The bootstrap," in: Heckman, J.J., and Leamer, E. (Eds.), *Handbook of Econometrics*, vol.5, ch. 52, 3159–3228, Elsevier.

[28] Horowitz, J. L. (2001b), "The Bootstrap and Hypothesis Tests in Econometrics," *Journal of Econometrics,* 100, 37–40.

[29] Kilian, L. (2007), Comment on "On the fit of new Keynesian models," by Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R., *Journal of Business and Economic Statistics,* 25,156–159.

[30] Le, V. P. M., Meenagh, D., Minford, P., and Wickens, M. R. (2010) "Two Orthogonal Continents? Testing a Two-country DSGE Model of the US and the EU Using Indirect Inference," *Open Economies Review*, 2010, vol. 21, issue 1, pages 23-44.

[31] Le, V. P. M., Meenagh, D., Minford, P., and Wickens, M. R. (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.

[32] Le, V. P. M., Meenagh, D. and Minford, P. (2012), "What causes banking crises? An empirical investigation," *Cardiff working paper* No E2012/14, Cardiff University, Cardiff Business School, Economics Section; also CEPR discussion paper no 9057, CEPR, London.

[33] Le, V. P. M., Minford, P., and Wickens, M. R. (2013), "A Monte Carlo procedure for checking identification in DSGE models," *Cardiff working paper* E2013/4, Cardiff Economics Working Papers, Cardiff University, Cardiff Business School, Economics Section; also CEPR discussion paper 9411.

[34] Le, V. P. M., Matthews, K., Meenagh, D., Minford, P., and Xiao, Z. (2014), "Banking and the Macroeconomy in China: A Banking Crisis Deferred?" *Open Economies Review*, vol. 25, issue 1, pages 123–161.

[35] Liu, C. and Minford, P. (2014a), "Comparing behavioural and rational expectations for the US post-war economy," *Economic Modelling*, vol. 43, issue C, pages 407–415.

[36] Liu, C. and Minford, P. (2014b), "How important is the credit channel? An empirical study of the US banking crisis," *Journal of Banking and Finance*, Volume 41, April, Pages 119–134.

[37] Lucas, R. E. (1976), "Econometric policy evaluation: A critique," Carnegie Rochester Conference Series on Public Policy No. 1, The Phillips Curve and Labour markets, K. Brunner and A. Meltzer, eds., supplement to *Journal of Monetary Economics.*

[38] McCallum, B. T. (1976), "Rational expectations and the natural rate hypothesis: some consistent estimates," *Econometrica,* 44, 4–52.

[39] Meenagh, D., Minford, P. and Wickens, M. R. (2009), "Testing a DSGE Model of the EU Using Indirect Inference," *Open Economies Review*, vol. 20, issue 4, pages 435–471.

[40] Meenagh, D., Minford, P., Nowell, E. and Sofat, P. (2010), "Can a real business cycle model without price and wage stickiness explain UK real exchange rate behaviour?" *Journal of International Money and Finance*, vol. 29, issue 6, pages 1131–1150.

[41] Meenagh, D., Minford, P. and Wickens M. R. (2012), "Testing macroeconomic models by indirect inference on unfiltered data," *Cardiff Working Paper* No E2012/17, Cardiff University, Cardiff Business School, Economics Section; also CEPR discussion paper no 9058, CEPR, London.

[42] Minford, P. and Ou, Z. (2013), "Taylor Rule or optimal timeless policy? Reconsidering the Fed's behavior since 1982," *Economic Modelling*, vol. 32, issue C, pages 113–123.

[43] Minford, P., Theodoridis, K. and Meenagh, D. (2009), "Testing a Model of the UK by the Method of Indirect Inference," *Open Economies Review*, vol. 20, issue 2, pages 265–291.

[44] Minford, P., Ou, Z. and Wickens, M. R. (2012), "Revisiting the Great Moderation: policy or luck?" *Cardiff working paper* E2012/9, Cardiff Economics Working Papers, Cardiff University, Cardiff Business School, Economics Section; forthcoming Open Economies Review.

[45] Sims, C. A. (1980), "Macroeconomics and reality," *Econometrica*, 48, 1–48.

[46] Sims, C. A. (2007), Comment on "On the fit of new Keynesian models," by Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R., *Journal of Business and Economic Statistics,* 25,152–154.

[47] Smets, F., and Wouters, R. (2003), "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area," *Journal of the European Economic Association*, 1(5), p1123–1175.

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

[49] Smith, A. (1993), "Estimating nonlinear time-series models using simulated vector autoregressions," *Journal of Applied Econometrics,* 8, S63–S84.

[50] Watson, M. (1993), "Measures of fit for calibrated models," *Journal of Political Economy* 101, 1011–1041.

[51] Wieland, V. and Wolters, M. H. (2012), "Forecasting and policy making," in G. Elliott and A. Timmerman (eds.), *Handbook of Economic Forecasting*, Vol. 2, Elsevier

[52] Wickens, M. R. (1982), "The efficient estimation of econometric models with rational expectations," *Review of Economic Studies,* 49, 55–67.

[53] Wickens, M. R. (2014), " How useful are DSGE macroeconomic models for forecasting?" *Open Economies Review*, 25, 171–193.