Open Econ Rev (2016) 27:1–38
DOI 10.1007/s11079-015-9377-5

RESEARCH ARTICLE

Testing Macro Models by Indirect Inference: A Survey for Users

Vo Phuong Mai Le[1] · **David Meenagh**[1] · **Patrick Minford**[1,2] · **Michael Wickens**[1,2,3] · **Yongdeng Xu**[1]

Published online: 25 September 2015

Abstract With Monte Carlo experiments on models in widespread use we examine the performance of indirect inference (II) tests of DSGE models in small samples. We compare these tests with ones based on direct inference (using the Likelihood Ratio, LR). 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 II test is substantially greater, both because the LR is applied after re-estimation of the model error processes and because the II test uses the false model's own restricted distribution for the auxiliary model's coefficients. This greater power allows users to focus this test more narrowly on features of interest, trading off power against tractability.

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. Programmes to implement the methods described in this paper can be downloaded freely and at no cost from www.patrickminford.net/Indirect.

✉ Patrick Minford
Patrick.minford@btinternet.com

Vo Phuong Mai Le
LeVP@cf.ac.uk

David Meenagh
Meenaghd@cf.ac.uk

Yongdeng Xu
Xuy16@cf.ac.uk

[1] Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

[2] CEPR 77 Bastwick St, London EC1V 3PZ, UK

[3] University of York, York YO10 5DD, UK

Keywords Bootstrap · DSGE · New Keynesian · New classical · Indirect inference · Wald statistic · Likelihood ratio

JEL Classification C12 · C32 · C52 · E1

1 Introduction

An unresolved issue in macroeconomics is how best to evaluate the empirical performance of DSGE models. 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 consider why we find this. We also show how this new testing procedure enables users such as policymakers to exploit the ability of the test and its associated estimator to focus on key features of macro behaviour; this allows them to find tractable models that are relevant to their purposes and then to discover whether these models can with total reliability evaluate the policy reforms they are interested in.

The paper is set out as follows. In Section 2 we consider how in recent work DSGE models have been evaluated 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 et al. (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 et al. (1999), again on both stationary and non-stationary data. In Section 6 we consider why the two testing methods have such different power, drawing on available asymptotic analysis as well as further Monte Carlo experiments. In Section 7 we show how the testing methods we propose can be used in practice to reduce model uncertainty for a user with a clear purpose such as policy reform. Our final section presents our conclusions.

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, 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 or mechanisms causing them to affect real variables; 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 and 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 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.

[1] In a recent 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).

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 ‘misspecified’. 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 (2013). 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 (2013). 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[2].

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

[2] In Appendix 1 we review some recent studies of macro models using this method.

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 et al. (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[3]. 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

[3] 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.

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 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.[4] In Le et al. (2011) we 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 et al. (2011). A particular concern with the bootstrap has been its consistency under conditions of near-unit roots.

[4] 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.

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 postwar 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.[5]

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.[6]

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

[5] 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 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 is FIML. Thus again the LR test cannot have power against a well-specified model that is freely estimated by FIML.

[6] 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 high.

Table 1 Rejection rates for wald and likelihood ratio for 3 variable VAR(1)

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

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 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 $-x$ %.[7]

Table 1 shows that the power of the Indirect Inference Wald test is substantially greater than that of the Direct Inference LR test. With 5 % mis-specification, the Wald statistic rejects 87 % of the time (at the 95 % confidence level) while the LR

[7] 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 re-estimation, 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 re-estimate 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.

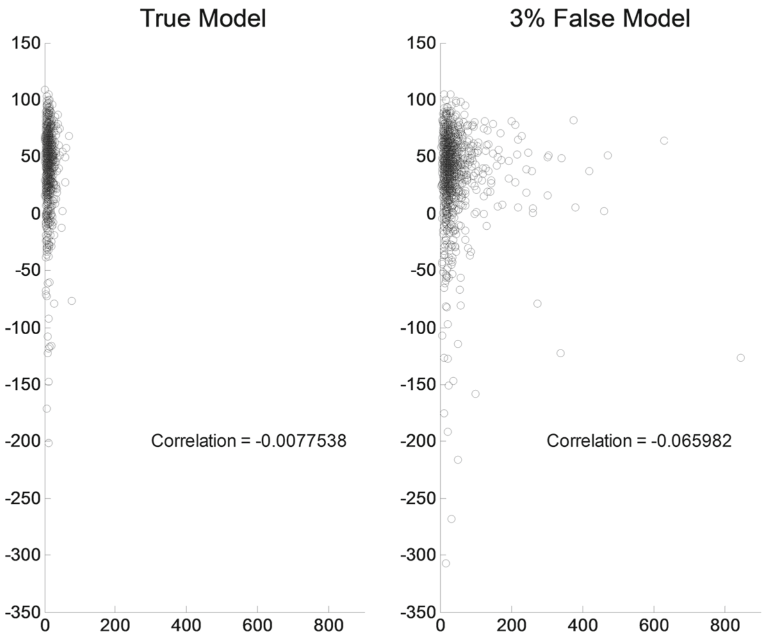


Fig. 1 Scatter plots of indirect inference (Wald; horizontal scale) v. Direct inference (log LR; vertical scale) for 1000 samples of true model (3 Variable VAR(1))

test rejects 13 % 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 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, Fig. 2, 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 behaviour.

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

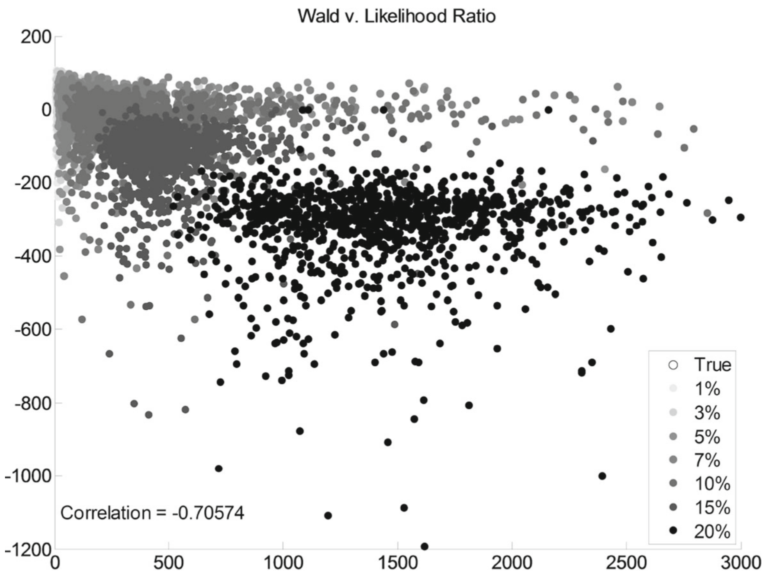


Fig. 2 Scatter plots of indirect inference (Wald; horizontal scale) v. Direct inference (log LR; vertical scale) for true and false models (some outliers taken out for clarity of scale)(3 Variable VAR(1))

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 power of direct inference based on a Likelihood Ratio test (using the LIML method

Table 2 Rejection Rates at 95 % level for varying VARs

Indirect inference								
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
Direct inference								
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

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.

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

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 3. 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.[8] 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 % competitive markets and a one-quarter

[8] 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.

Table 3 Small Sample Estimation Bias Comparison (II v. LR)

			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 Power of the test to reject a false model

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

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.

The results in Table 4 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 Eqs. 6 or 7 can act as the auxiliary model. Here we focus on Eq. 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 (7) in one stage by OLS. Using Monte Carlo experiments, Meenagh et al. (2012) show that this procedure is

Table 5 Rejection rates at 95 % level for varying VARs (non-stationary data)

Indirect inference								
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
Direct inference								
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

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 et al. (2012). The rejection probabilities for the Wald and LR tests are reported respectively in Table 5. Once more the test based on indirect inference has far more power than the direct LR test.

5.2 Extension to the 3-equation New Keynesian Model

The results for the 3-equation New Keynesian inflation model are reported for stationary data in Table 6 and for non-stationary data in Table 7 . 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 fasthan 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.

Table 6 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	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

6 Why Does the Indirect Inference Test have Greater Power than the Likelihood Ratio Test?

What we have shown so far is that in small samples the direct inference LR test has far less power than the Indirect Inference Wald test. The LR test is familiar. Let us review exactly the way the Indirect Inference test is carried out. Notice that we simulate the DSGE model to find its implied distribution for the VAR coefficients; the Wald test then checks whether the data-estimated VAR coefficients lie within the 95 % bounds of this distribution - i.e. whether the DSGE-model-restricted distribution 'covers' the data-based VAR coefficients at the specified significance level. However, we could have done the test differently, in effect 'the other way round', creating the distribution of the data-estimated VAR coefficients and asking

Table 7 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

whether this data-based distribution covers the DSGE-model-restricted VAR coefficients (which we can obtain as the mean of the model-implied VAR coefficients distribution). This is the way in which a standard classical Wald test is performed: thus the data-based distribution (which comes from the true model, unrestricted) is treated as the null and the alternative hypothesis is tested against it in this way. This unrestricted Wald is a transformation of the LR test - as is familiar from standard econometrics. We can also obtain it by bootstrapping the estimated VAR. This distribution is unrestricted because it uses the estimated VAR without imposing on it the restrictions of the true (but unknown) model. Thus when bootstrapping the estimated VAR one holds the α_T constant, merely bootstrapping the VAR errors (which are linear combinations of the structural errors); whereas if one bootstrapped the true structural model, one would be capturing the overall variation in α_S across samples due to both the errors and their interaction with the structure of the DSGE model[9]. It turns out that this is an important distinction between the two Walds. We will see below that the Wald using the restricted distribution - the 'restricted Wald' - creates a more powerful test than the one based on the unrestricted distribution - the 'unrestricted Wald'. For now we will simply explore the theoretical differences between the restricted Wald on the one hand and the LR statistic or the unrestricted Wald on the other.

Meenagh et al. (2015), whom we follow closely in this section, show that the three tests are asymptotically equivalent when the DSGE model being tested is true. However when the DSGE model is false the restricted Wald test is not asymptotically equivalent to the other two. By using the distribution of the model-restricted VAR coefficients it generates increased precision of the variance matrix of the coefficients of the auxiliary model and so improves the power of the Wald test.

[9] This can be seen formally by noting that the α coefficients re-estimated from the ith bootstrap of the unrestricted VAR (found from the T data sample) are:
$\widehat{\alpha}_i^{UNR} = f_{OLS}\{\widehat{y}_i^{UNR} = \widehat{y}_i^{UNR}[\widehat{\alpha}_T(\theta, \epsilon_T), \eta_i]\}$
where θ is the vector of structural model coefficients (including those of the error processes), ϵ the vector of structural innovations, η that of VAR innovations, f_{OLS} is the OLS estimator function to obtain the α from the y.

Now compare the analogous estimates with restricted VAR bootstraps:
$\widehat{\alpha}_i^{RES} = f_{OLS}\{\widehat{y}_i^{RES} = \widehat{y}_i^{RES}[\theta, \epsilon_i] = \widehat{y}_i^{RES}[\widehat{\alpha}_i(\theta, \epsilon_i), \eta_i]\}$
We can see that these α OLS estimates come from y simulated directly from the structural model and that these in turn have a VAR representation consisting of two elements, the direct effect of η as before plus the indirect effect of ϵ, θ on α. It is this last extra element that creates the rich variation in resampled data behaviour reflecting the DSGE model's structure interacting with the structural errors.

In terms of the example discussed below in the text where we consider the own-persistence VAR parameters of inflation and interest rates, what is happening is that with restricted bootstraps model-simulated samples in which inflation is not persistent will typically also be those where interest rates are also not persistent, and vice versa, because the model implies a strong connection between the two variables; thus estimated covariation in these own-persistence VAR parameters ($\widehat{\alpha}_i(\theta, \epsilon_i)$) will show up in the resampled data. With unrestricted bootstraps this covariation is not included; instead the VAR parameters generating the data are held constant at those in the data sample, $\widehat{\alpha}_T(\theta, \epsilon_T)$. Notice that the variation due to the direct effect of the innovations, η_i, is the same in both cases

6.1 Summary: Why the Power is Different

With these introductory remarks we are now in a position to analyse the reasons for the difference in power we have found between the two small sample tests, LR and our Indirect Inference Wald, IIW. In summary we find two main reasons: a) they are carried out with different procedures; b) even when the same procedures are followed, the two tests differ in power by construction. Let us now discuss these in turn.

6.1.1 Reason a): The Tests Employ Different Procedures so the Comparison is of Different Models

We have seen above how when re-estimation is permitted using the LR test, power is reduced. Thus when one is finding the rejection rate when parameter values are falsified, we saw that with the LR test the re-estimation of the error process to bring the model back on track reduced the rejection rate. This can be illustrated by comparing the power of the LR test in which the autoregression coefficients are re-estimated, as above, with an LR test in which the degree of falsification of the autoregressive coefficients is pre-specified, as for the Wald test above. We employ a 3-equation NK model for the comparison. As expected, the results in Table 8 below shows that the LR test with pre-specified autoregressive coefficients has considerably greater power than the test using re-estimated autoregressive coefficients.

We further found that the power of the LR test against a completely mis-specified model was virtually nil, because the FIML estimator of the mis-specified model manages to 'data mine' highly effectively in fitting the wrong model - see Table 8 below. The point here is that the power is again eliminated by bringing the model, across all its parameters and not merely the AR ones, onto track with the data.

6.1.2 Reason b) Comparative Power when the LR and Indirect Inference Wald Procedures are Like-for-like

In the above comparison of the joint distribution of the two coefficients of interest, the data simulated from the structural model gave serially correlated structural error processes. In order to make the estimates of their joint distribution compatible with the original Smets-Wouters estimation strategy, first-order autoregressive processes were fitted to these structural errors for each bootstrap sample. In calculating the power of the tests we proceed a little differently in order that the tests are based on the same assumptions when the structural model is falsified. We now fix both θ (the vector of structural coefficients of the DSGE model) and ρ (the vector of coefficients of the autoregressive error processes). Each is falsified by x %. We do not, however, falsify the innovations, maintaining them as having the original true distribution. This is a matter of convenience as we could extract the exact implied false error innovations, as implied by each data sample, θ and ρ. But this extraction is a long and computationally-intensive process requiring substantial iteration (because the model expectations depend on the errors while the errors in turn depend on the expectations). We simply assume, therefore, that the model is false in all respects except for the innovations. For our purposes here, which is to determine the relative

Table 8 Comparing power due to wrong parameter values

3-equation NK model (no lags)

Rejection rate of false models at 95 % confidence: T=200

	Re-estimated $\rho's$	Pre-specified $\rho's$
True	5.0	5.0
1 %	5.0	5.0
3 %	5.3	9.6
5 %	6.1	20.2
7 %	8.0	39.1
10 %	15.4	63.7
15 %	48.1	90.7
20 %	75.6	98.9

power of the two tests when faced with exactly the same falsified models, this creates no problems. We use the SW model as the true model with a sample size of 200 throughout. Our findings are reported in Table 9.

We find that the two test statistics, LR and Wald, generate similar power when the unrestricted Wald test is used, i.e. based on the observed data (the unrestricted VAR). This is what we would expect since the unrestricted Wald, as we have seen, is simply a transformation of the LR test. Focusing on the main case, which is a 3VAR1, and taking 5 % falseness as our basic comparison, we see that the rejection rate for the LR test is 38 %. For the unrestricted Wald test, based on the unrestricted VAR, the rejection rate is 31 %. However, using the restricted Wald (IIW) test the power rises to 71 %, nearly double that of the two other tests.[10]

Understanding the Extra Power Provided by using the Restricted rather than the Unrestricted Wald Tests In our numerical comparison of the two tests our structural model is the Smets and Wouters model (2007). This is a DSGE model which has a high degree of over-identification (as established by Le et al. (2013)). It has 12 structural parameters and 8 parameters in the error processes. It implies a reduced-form VAR of order 4 with seven observable endogenous variables, i.e. a 7VAR4, (Wright, 2015). This has 196 coefficients. The size of the VAR in a IIW test and the number of variables is usually lower than a 7VAR4.

We concentrate on the dynamic response to own shocks of inflation and the short-term nominal interest rate. We focus on the three variables of the above New Keynesian model: inflation, the output gap and the nominal interest rate. We use a 3VAR1 in these variables as the auxiliary model. We then examine the own-lag coefficients for inflation and the short-term interest rate.

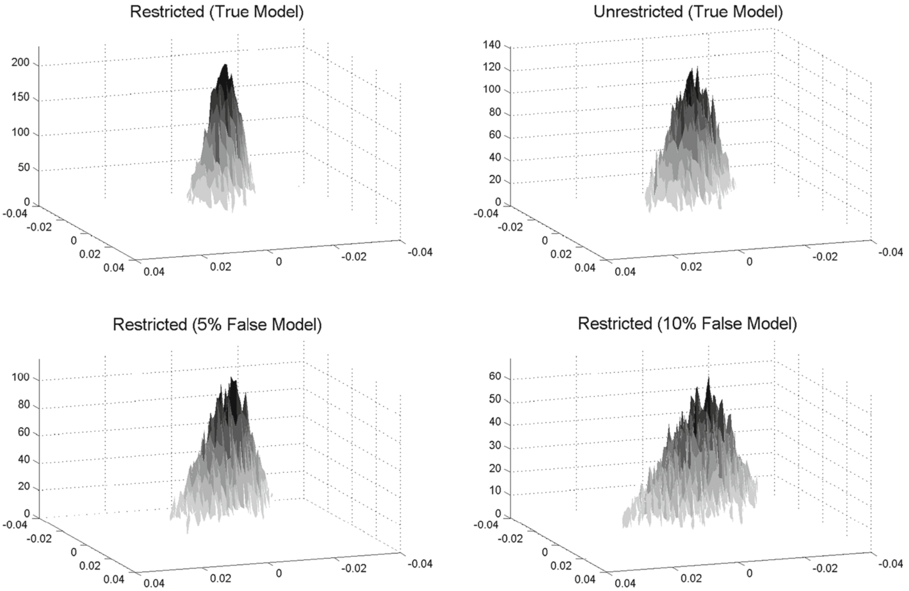
[10] The unrestricted Wald test uses the variance matrix of the auxiliary model. When the VAR has a very large number of coefficients the variance matrix of the coefficients has a tendency to become unstable; this occurs even when the number of bootstraps is raised massively (e.g. to 10000). This is due to over-fitting in small samples (here the sample size is 200); there is then insufficient information to measure the variance matrix of the VAR coefficients.

Table 9 Comparison of rejection rates at 95% level for Indirect Inference and Direct Inference

VAR — no of coeffs WALD TEST with unrestricted VAR	TRUE	1 %	3 %	5 %	7 %	10 %	15 %	20 %
2 variable VAR(1) — 4	5.0	6.2	20.3	69.6	61.0	99.8	100.0	100.0
3 variable VAR(1) — 9	5.0	3.4	7.5	30.7	75.0	97.4	100.0	100.0
3 variable VAR(2) — 18	5.0	3.8	5.2	19.1	57.5	84.3	98.4	99.5
3 variable VAR(3) — 27	5.0	3.9	6.4	21.6	54.5	84.0	97.5	98.7
5 variable VAR(1) — 25	5.0	2.8	3.2	2.6	5.4	6.2	4.5	100.0
7 variable VAR(3) — 147	5.0	5.1	3.4	1.4	0.9	0.2	0.0	100.0
IIWALD test (with restricted VAR)								
2 variable VAR(1) — 4	5.0	9.8	37.7	80.8	96.8	100.0	100.0	100.0
3 variable VAR(1) — 9	5.0	9.5	36.1	71.0	98.1	100.0	100.0	100.0
3 variable VAR(2) — 18	5.0	8.3	35.5	80.9	96.9	100.0	100.0	100.0
3 variable VAR(3) — 27	5.0	9.2	32.9	78.0	95.1	100.0	100.0	100.0
5 variable VAR(1) — 25	5.0	17.8	85.5	99.8	100.0	100.0	100.0	100.0
7 variable VAR(3) — 147	5.0	77.6	99.2	100.0	100.0	100.0	100.0	100.0
Likelihood Ratio test								
2 variable VAR(1) — 4	5.0	12.0	28.3	45.9	63.4	83.2	97.0	99.7
3 variable VAR(1) — 9	5.0	9.4	21.8	37.5	58.9	84.0	99.0	100.0
3 variable VAR(2) — 18	5.0	8.9	20.7	36.8	57.6	82.9	98.7	100.0
3 variable VAR(3) — 27	5.0	8.9	20.4	36.7	56.7	82.2	98.7	100.0
5 variable VAR(1) — 25	5.0	8.9	22.4	44.3	68.6	89.6	99.6	100.0
7 variable VAR(3) — 147	5.0	5.7	10.6	23.6	46.3	83.2	99.6	100.0

We estimate the coefficients of the 3VAR1 using the observed data for these three variables. We then find the distribution of the estimates of the two coefficients of interest by bootstrapping the VAR innovations. Next, we estimate the 3VAR1 using data for these three variables obtained by simulating the full SW model. The distribution of these estimates of the two coefficients is obtained by bootstrapping the structural innovations generating that sample.

Figure 3 displays the joint distributions of the two VAR coefficients based on 1) the observed data (the unrestricted VAR), 2) simulated data from the original estimates of the structural model (the restricted VAR), and 3) false specifications of the structural models by 5 % and 10 % (the 5 % false and 10 % false restricted VARs). One can see clearly that 2), the joint distribution based on simulated data from the original structural model, is both more concentrated and more elliptical (implying a higher correlation between the coefficients) than 1), that using the observed data. Increasing the falseness of the model causes 3), the joint distributions from the 5 % and 10 % false DSGE model, to become a little more dispersed and more elliptical; they are also located slightly differently but this is not shown as the distribution is centred on zero in all cases.



Restricted (True Model)

Unrestricted (True Model)

Restricted (5% False Model)

Restricted (10% False Model)

Fig. 3 Restricted VAR and unrestricted VAR coefficient distributions

Figure 4 shows how this affects the power of the Wald test for a model that is 5% false. The dot on the right of the Figure is the mean of the distribution. The test of this false model can be carried out in two ways. We have drawn the diagram as if the

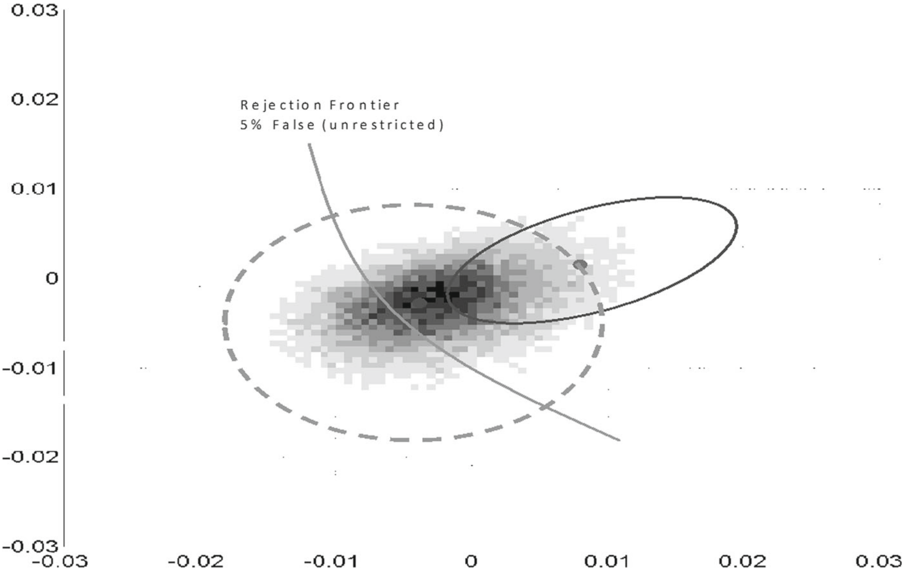


Fig. 4 Model- Curve to the left=Unrestricted; Ellipse to the right=Restricted

joint test of two VAR coefficients chosen have the same power as the overall test of all VAR coefficients.

The first way is to use the unrestricted Wald, using the observed data to estimate a 3VAR1 representation and to derive the joint distribution of the two coefficients by bootstrapping. The 5 % contour of such a bootstrap distribution is given by the dashed (close to circular) line; the thick curve to the left of the figure shows the critical frontier at which the 5 % false model is just rejected.

The second way is to use the restricted Wald, using the distribution implied by the simulated data. The ellipse to the right of the figure shows the 5 % contour of the resulting joint distribution. The results show that the second method has nearly double the power of the first. (Increasing the degree of falseness to 10% raises the power of both to 100 %.)

We can also look at Fig. 5 to see how the rotation of the ellipse due to the changing covariance of the two VAR coefficients can raise the power of the restricted Wald test. As the ellipse rotates, it covers less and less of the True model sample points. Thus not just the distance of the model's mean VAR coefficients from the True mean of the data-based ones but also the shape of the model's distribution for these coefficients and its rotation (both due to the model-implied covariance between the coefficients) with rising falseness determine the power of the test - i.e. how many of the data sample points it fails to cover. With the standard unrestricted Wald test the shape and rotation is fixed regardless of Falseness - one is always using the same distribution based on the data sample - and so only the distance varies with Falseness.

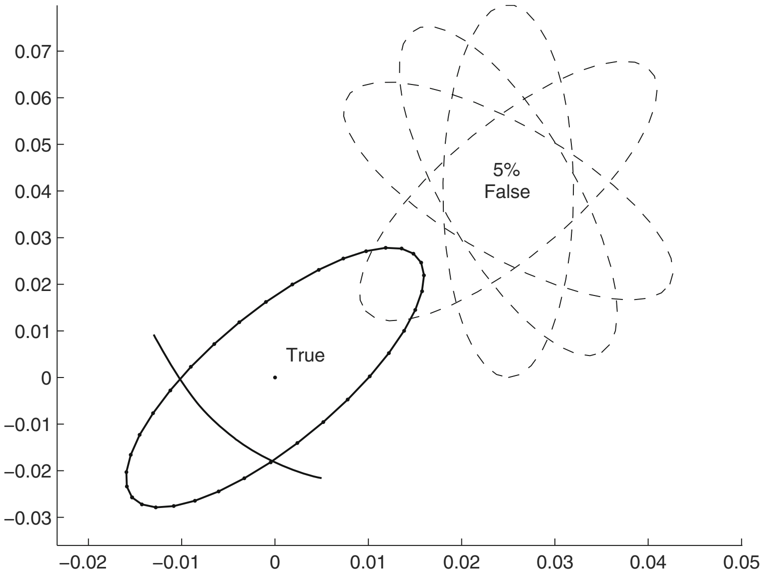


Fig. 5 Joint distribution of VAR coefficients rotates with changing false DSGE parameters

Table 10 Comparing power due to VAR order (3-equation NK model with no lags)

3-equation NK model — no lags (VAR(1) reduced form)
Rejection rates at 95 % confidence: T=200

	3 variable VAR(1)	3 variable VAR(2)
True	5.0	5.0
1 %	4.9	4.3
3 %	7.3	7.1
5 %	16.1	21.7
7 %	37.0	40.3
10 %	73.3	76.3
15 %	99.4	99.8
20 %	100.0	100.0

Exploiting the Extra Power of the Wald-type Test with DSGE-model-restricted Variance Matrix Thus when we eliminate the difference in procedures and test like-for-like we found the two tests are reasonably comparable in power when the indirect inference test is performed using the unrestricted Wald test which uses the variance of the unrestricted VAR (auxiliary) model. This turns out to be because the tests are approximately equivalent on a like-for-like basis. However, we showed above that extra power is delivered by the IIW test set out here, under which the DSGE model being tested is treated as the null hypothesis: in this case the Wald statistic uses the variance restricted by the DSGE model under test. This gives this restricted Wald test still greater power. (Table 10)

It may be possible to raise the power of the Wald test further. We suggest two ways this might be achieved:

1) extending the Wald test to include elements of the variance matrix of the coefficients of the auxiliary model;
2) including more of the structural model's variables in the VAR, increasing the order of the VAR, or both.

The basic idea here is to extend the features of the structural model that the auxiliary model seeks to match. The former is likely to increase the power of the restricted Wald test, but not the LR test, as this last can only ask whether the DSGE model is forecasting sufficiently accurately; including more variables is likely to increase the power of both. There is, of course, a limit to the number of features of the DSGE model that can be included in the test. If, for example, we employ the full model then we run into the objection raised by Lucas and Prescott against tests of DSGE models that "too many good models are being rejected by the data". The point is that the model may offer a good explanation of features of interest but not of other features of less interest, and it is the latter that results in the rejection of the model by conventional hypothesis tests. Focusing on particular features is a major strength of the Wald test.

Table 11 Comparing power due to VAR order (3-equation NK model with indexing lag)

3-equation NK model — with lag (VAR(2) reduced form)

Rejection rates at 95% confidence: T=200

	3 variable VAR(1)	3 variable VAR(2)
True	5.0	5.0
1 %	10.6	6.0
3 %	20.7	19.5
5 %	47.5	57.9
7 %	65.6	91.2
10 %	89.6	100.0
15 %	98.8	100.0
20 %	99.9	100.0

Consider now including an indexing lag in the Phillips Curve. This increases the number of structural parameters to 9 and the reduced-form solution is a VAR(2). The power of the Wald test is reported in Table 11 . Increasing the number of lags in the auxiliary model has clearly raised the power of the test.

This additional power is related to the identification of the structural model. The more over-identified the model, the greater the power of the test. Adding an indexation lag has increased the number of over-identifying restrictions exploitable by the reduced form. A DSGE model that is under-identified would produce the same reduced-form solution for different values of the unidentified parameters and would, therefore have zero power for tests involving these parameters.

In practice, most DSGE models will be over-identified - see Le et al. (2013). In particular, the SW model is highly over-identified. The reduced form of the SW model is approximately a 7VAR(4) which has 196 coefficients. Depending on the version used, the SW model has around 15 (estimatable) structural parameters and around 10 ARMA parameters. The 196 coefficients of the VAR are all non-linear functions of the 25 model parameters, indicating a high degree of over-identification.

The over-identifying restrictions may also affect the variance matrix of the reduced-form errors. If true, these extra restrictions may be expected to produce more precise estimates of the coefficients of the auxiliary model and thereby increase its power. It also suggests that the power of the test may be further increased by using these variance restrictions to provide further features to be included in the test.

7 Using These Methods to Test a Model

In this final section we discuss the results we have found in using the Smets-Wouters model for monetary and fiscal policy purposes in the context of the recent crisis and its aftermath. This work is all on US data for the period since the mid-1980s; we have not found it possible to mimic US behaviour for earlier data, we think because there has been substantial regime change before then - Le et al. (2014).

We start from the position that the model has credible micro-foundations but that we are searching for a variant of it that a) can allow for a banking system with the monetary base (M0) as an input into it b) can integrate the zero bound on the risk-free interest rate and Quantitative Easing together with bank regulation as policy tools; and c) can explain the behaviour of the three key macro variables: output, inflation and interest rates. This is because we want to find a model within which we can reliably explore policies that would improve these variables' behaviour, especially their crisis behaviour. There is of course a large macro literature in which claims are made for the efficacy of a variety of policy prescriptions; but here we just focus on the set of policies investigated for this model, to illustrate the power of our methods.

We will discuss the model's properties with these policies in a moment. But first let us note that we can test it two ways - by a Likelihood Ratio test for three key macro variables, inflation, output and interest rates and also by an IIW test on the same three variables. We choose these because they are focused on the behaviour of the three variables of interest to us as policymakers. The LR test measures how close the model gets to the data - essentially a forecasting test; notice at once that this is not really our interest but we are using it as a general specification test. It turns out that the LR test is not sensitive, at least for the SW model, to what variables are included in the test, no doubt becase if a model forecasts some variables well, it must be forecasting the other variables well that are closely linked to them. We carry out the LR test in the usual way, allowing the ρs to be re-estimated on the error processes extracted by LIML. The IIW test looks at how close the model gets to these three variables' data behaviour - which we are deeply interested in matching and represented by a VECM (which we rewrite as a VARX) here as the data is non-stationary. Thus with the IIW test we have carefully chosen its focus to match our policy interests; we could have chosen a broader group of variables which would have raised the test power but at the cost of possibly not finding a model that would fit their broader behaviour. Thus we see here that the focus of the test is a crucial aspect of the IIW test.

We now reproduce some Monte Carlo experiments for the SW model from Tables 1 and 5 above:

The basic point we want to emphasise from this comparison is that if this model passes the IIW test, we can be sure it is less than 7 % False whereas if it passes the LR test we can only be sure it is less than 15 % False under stationarised data; under non-stationary data, the relevant case here, we cannot even be sure it is less than 20 % False - in fact we find that it requires the model to be as much as 50 % False for it to be rejected roughly 100 % of the time.

When we now apply the two tests to the Monetary model discussed above, it passes both tests. We can now compare how our policy analysis would vary with the two test approaches. (Table 12)

Our basic policy results when we treat the model as True are summarised in the first row of the following Table 13:

If we use the IIW test we know that our model could be up to 7 % False but no more. We can discover the effect of this degree of Falseness on our policy results by

Table 12 Rejection Rates for Wald and Likelihood Ratio for 3 Variable VAR(1)

Percent Mis-specified	Wald	LR	Wald	LR
	Stationary data		Non-stationary data	
True	5.0	5.0	5.0	5.0
1	19.8	6.3	7.9	5.2
3	52.1	8.8	49.2	5.8
5	87.3	13.1	97.8	6.2
7	99.4	21.6	100.0	7.4
10	100.0	53.4	100.0	9.6
15	100.0	99.3	100.0	15.6
20	100.0	99.7	100.0	26.5

redoing the whole policy exercise with the parameters disturbed by 7 %. We obtain the results shown in the second row of Table 13.

In investigating the power of the test, we have simply assumed that we are presented with a False set of parameters somehow from the estimation process. We can then ask what power can we have against a quite mis-specified model whose parameters are simply different. We have looked at this for the model here, by asking what the power is against a quite different model - say a New Classical model versus as assumed True SW model. The power is 100 %; it is always rejected. So we can be quite sure the True model is not something quite different.

Between these two things we therefore have a lot of reassurance. First, if the model is not well-specified, it will be certainly rejected. Second, if the model is well-specified, then models up to 7 % distant from it could be True; and our policy conclusions can be tested for robustness within this range as we have done here.

Table 13 Policy analysis when models have varying falseness

Frequency of crisis (expected crises per 1000 years)	Base case	Monetary Reform	PLT	NGDPT	PLT+ Mon.Reform	NGDPT+ Mon.Reform
Policy exercise						
when model is True	20.8	6.62	2.15	1.83	1.41	1.31
when model is 7 % False	57.4	18.6	10.3	8.7	11.8	10.3
when model is 15 % False	63.6	*Explosive*	19.4	19.6	19.4	17.4
when model is 50 % False	70.4	*Explosive*	33.3	33.4	34.4	34.2

Notes:

Base Case: monetary policies as estimated over the sample period

Monetary Reform: running a Monetary Base rule targeted on the credit premium side by side with a Taylor Rule

PLT:substituting Price Level Target for Inflation Target in Taylor Rule

NGDPT: substituting Nominal GDP target for inflation and output targets in Taylor Rule

If we use the LR test we know the model could be up to 50 % False - we cannot guarantee to reject a model that is less false than this. For example a 15 % False model will be rejected only a third of the time. If we now redo the exercise for a 15 % disturbance to the parameters we obtain the third row of Table 13. Now our policy is plainly vulnerable. The frequency of crises under the current regime goes up to once every 15 years; with NGDPT+monetary reform it only comes down to once every 50-60 years. This is on the borderline of acceptability.

If we look at the 50 % false case, shown in the last row of Table 13, it is disastrous. First, only just under half of the bootstrap simulations have sensible solutions. If we take those that do, we can see that the prevalence of crises under the existing regime would be much greater, at one every 14 years. As with 15 % False the monetary reform regime is explosive. The other regimes all generate crisis frequency of around one every 30 years which is far from acceptable.

To make matters worse, we have seen that the LR test has virtually no power against model misspecification, so that we cannot be sure that a misspecified model with yet other, possibly even worse, results might be at work.

What this is showing us is that according to the LR test versions of our model that could be true imply much higher frequency of crises than in the estimated case and the monetary policy regimes suggested as improvements could either give explosive results or produce an improvement in the crisis frequency that is quite inadequate for policy purposes. In other words the policymaker cannot rely on the model policy results. But using the IIW test we can be sure that the recommended policies will deliver the results we claim.

7.1 Can Estimation Protect us Against Falseness?

But would this vulnerability not be reduced if we take ML estimation seriously? Unfortunately, as we saw above, estimation by ML gives us no guarantees of getting close to the true parameters. It is well-known to be a highly biased estimator in small samples - with an average absolute estimation bias across all parameters of nearly 9% in our Monte Carlo experiment above (see Table 3). Bearing in mind that our 'falseness' measure assumes x as the absolute bias, alternating plus and minus, this suggests that FIML will on average give us this degree of falseness; in any particular sample it could be much larger therefore.

We also looked above at whether the Indirect Inference estimator could give us any guarantees in this respect. This estimator was much less biased in small samples, with an average absolute bias about half that of FIML, as again shown in Table 3. However, again this can give us no guarantees of the accuracy of the estimates in any particular sample.

It follows that we are essentially reliant on the power of the test, in the sense that this can guarantee that our model is both well specified and no more than 7 % false under indirect inference, because if it were either it would have been rejected with complete certainty.

The dimension in which we have carried out this examination of the model's reliability in the face of what we might call 'general falseness'. It may be also that the model's performance is sensitive to the values of one or two particular parameters

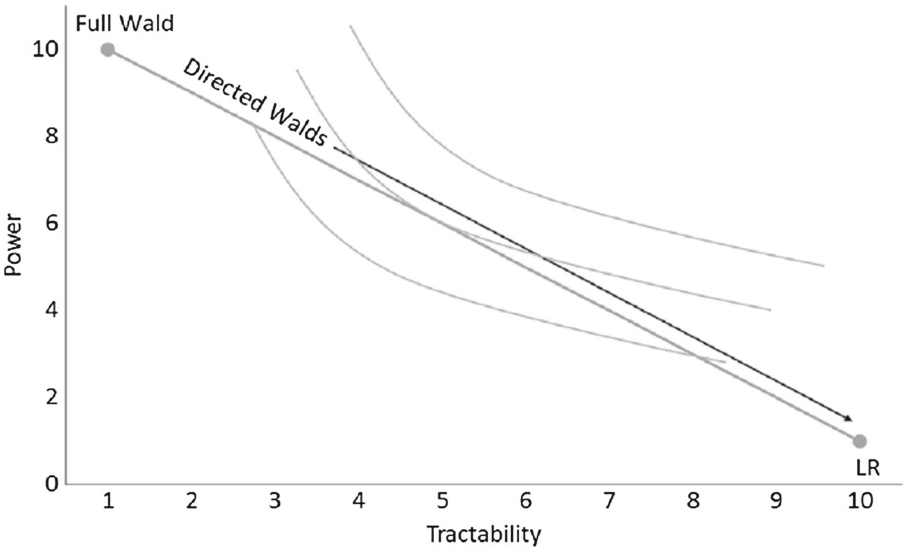


Fig. 6 Maximising Friedman utility

and if so we would also need to focus on the extent to which these might be false, how far the test's power can protect us against this and how sensitive the model is within this range. This further investigation can be carried out in essentially the same way as the one we have illustrated with general falseness.[11]

7.2 Choosing the Testing Procedure

Thus what we have illustrated in this section is how macro models can be estimated and tested by a user with a particular purpose in mind. The dilemma a user faces is the trade-off between test power (i.e. the robustness to being false of a model that marginally passes the test) and model tractability (i.e. the relevance for the facts to be explained of a model that marginally passes the test). Different testing procedures give different trade-offs as we have seen and is illustrated in Fig. 6. Thus the Full Wald test gives the greatest power; but a model that passes this test will have to reflect the full complexity of detailed behaviour and thus be highly intractable. At the other extreme the LR test is easy to pass for a simple and tractable model; but it has very low power. In between lie Wald statistics with increasing 'narrowness' of focus as we move away from the Full Wald. These offer lower power in return for higher tractability - somewhere along their trade-off will be chosen by the policymaker, as shown in Fig. 6.

[11] The LR test and the Monte Carlo results for power are based on various versions of the Smets-Wouters model and varying data samples. Our aim is to illustrate the method of policy analysis. Ideally the policymaker should redo all this work on the model and data sample being used.

In order for us to find a tractable model we have to allow a degree of falseness in the model with respect to the data features other than those the policymaker prizes. The way to do this is to choose an Indirect Inference test that focuses tightly (in a 'directed' way) on the features of the data that are relevant to our modelling purposes.

To apply these methods it is necessary to a) estimate and test the model, b) assess which 'directed' test to choose, c) assess the power in the case of the model being used. We have programmes to do these things which we are making available freely to users - Appendix 2 shows the steps involved in finding the Wald statistic, as carried out in these programmes[12].

8 Conclusions

In this paper we examine the workings of the Indirect Inference Wald test of a macroeconomic model, typically a DSGE model. We show how the model can be estimated by Indirect Inference and how much power the test has in small samples against falseness in the estimated parameters as well as against complete model misspecification. We perform numerous Monte Carlo experiments with widely-used DSGE models to establish the extent of this power. We consider how the test can be focused narrowly (via a 'directed Wald') on features of the model in which the user is interested, echoing Friedman's advice that models should be tested 'as if true' according to their ability to explain features of the data the user is concerned about. For a user of a model with a clear purpose, for example a monetary policymaker, this testing method offers an attractive trade-off between the chances of finding a model to pass the test and the power of the test to reject false models. Thus the user can determine whether the model found can be assumed to be reliable enough to use for a policy exercise, by seeing whether it is robust to the potential degree of falseness it could be open to. In this way users can discover whether their models are 'good enough', in Friedman's original sense, for the purposes intended, and the model uncertainty facing them can be reduced and even eliminated. Tailor-made programmes to carry out this procedure are now available to applied macro economists.

We benchmarked the IIW test against the widely-used Likelihood Ratio (LR) test, test. A key finding is that, in small samples, tests based on the IIW test have much greater power than those based on the LR test. This finding is at first sight puzzling as the LR test can be transformed into a standard Wald test, which in turn can be obtained by indirect inference using the unrestricted variance matrix of the auxiliary model coefficients estimated on the data. We attempted to explain why this result occurs.

We find that the difference in power in small samples of the two tests can be attributed to two things. First, for the LR test the autoregressive processes of the structural errors are normally re-estimated when carrying out the test. This 'brings

[12] Programmes to implement the methods described in this paper can be downloaded freely and at no cost from http://www.patrickminford.net/Indirect.

the model back on track' and as a result undermines the power of this test as it is, in effect, based on the relative accuracy of one-step ahead forecasts compared with those obtained from an auxiliary VAR model.

Second, additional power of the IIW test arises from its use of the restricted variance matrix of the auxiliary model's coefficients, determined from data simulated using the restrictions on the DSGE model. These may give both give more precise estimates of these coefficients and provide further features of the model to test. The greater the degree of over-identification of the DSGE model, the stronger this effect. This suggests that for a complex, highly restricted, model like that of Smets and Wouters, the power of the Indirect Inference Wald test can be made very high even in small samples. Because a test of all of the properties of a DSGE model is likely to lead to its rejection, it is preferable to focus on particular features of the model and their implications for the data. This is where the IIW test can flexibly be tailored to optimise the ratio of power to tractability.

In sum, we find that the IIW test can become a formidable weapon in the armoury of the users of macro models, enabling them to estimate a model that can pass the test when suitably focused and then to check its reliability in use against such potential inaccuracy as cannot be ruled out by the power of the test.

Appendix 1: Available IIW tests of macro models

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 14.

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

Table 14 Tests using varying VARs

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

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 14) both tests give consistent results.

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 15 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 three columns denote the country, sample episode and the model studied; the fifth column provides the name of the authors and the reference. The fourth 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 etal. (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 et al. (1999) passes the test after re-estimation and can even match higher-order VARs - see Minford and Ou (2013), Liu and Minford (2014a, b), and also Minford et al. (2012) 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.

Table 15 Summary of recent tests of DSGE models

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, π, 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, π, R /98.7	Le et al. (2011)
US	1984–2004	Smets-Wouters hybrid[⊤]	Indirect estimation	y, π, R /83.8	Le et al. (2011)
US	1981–2010	3-eqn NK(Rational Exp.)	Indirect estimation	y, π, 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, π, R/45.4	Liu and Minford (2014b)
China	1978–2007	Smets-Wouters hybrid[⊤]	Indirect estimation	y, π, R /69.0	Dai et al. (2014)
China	1991–2011	Smets-Wouters hybrid+bkg[⊤‡*]	Indirect estimation	y, π, 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

Appendix 2: Steps in deriving the Wald statistic

The following steps summarise our implementation of the Wald test by bootstrapping:

Step 1: Estimate the errors of the economic model conditional on the observed data and θ_0.

Estimate the structural errors ε_t of the DSGE macroeconomic model, $x_t(\theta_0)$, given the stated values θ_0 and the observed data. The number of independent structural errors is taken to be less than or equal to the number of endogenous variables. The errors are not assumed to be normally distributed. Where the equations contain no expectations the errors can simply be backed out of the equation and the data. Where there are expectations estimation is required for the expectations; here we carry this out using the robust instrumental variables methods of McCallum (1976) and Wickens (1982), with the lagged endogenous data as instruments — thus effectively we use the auxiliary model VAR. An alternative method for expectations estimation is the 'exact' method; here we use the model itself to project the expectations and because these depend on the extracted residuals there is iteration between the two elements until convergence.

Step 2: Derive the simulated data

On the null hypothesis the $\{\varepsilon_t\}_{t=1}^{T}$ are the structural errors. The simulated disturbances are drawn from these errors. In some DSGE models, including the SW model, many of the structural errors are assumed to be generated by autoregressive processes rather than being serially independent. If they are, then under our method we need to estimate them. We derive the simulated data by drawing the bootstrapped disturbances by time vector to preserve any simultaneity between them, and solving the resulting model using Dynare (Juillard 2001). To obtain the N bootstrapped simulations we repeat this, drawing each sample independently. We set $N = 1000$.

Step 3: Compute the Wald statistic

We estimate the auxiliary model — a VAR(1) — using both the actual data and the N samples of simulated data to obtain estimates a_T and $a_S(\theta_0)$ of the vector α. The distribution of $a_T - \overline{a_S(\theta_0)}$ and its covariance matrix $W(\theta_0)^{-1}$ are estimated by bootstrapping $a_S(\theta_0)$. The bootstrapping proceeds by drawing N bootstrap samples of the structural model, and estimating the auxiliary VAR on each, thus obtaining N values of $a_S(\theta_0)$; we obtain the covariance of the simulated variables directly from the bootstrap samples. The resulting set of a_k vectors $(k = 1, ..., N)$ represents the sampling variation implied by the structural model from which estimates of its mean, covariance matrix and confidence bounds may be calculated directly. Thus, the estimate of $W(\theta_0)^{-1}$ is

$$\frac{1}{N}\Sigma_{k=1}^{N}(a_k - \overline{a_k})'(a_k - \overline{a_k})$$

where $\overline{a_k} = \frac{1}{N}\Sigma_{k=1}^{N} a_k$. We then calculate the Wald statistic for the data sample; we estimate the bootstrap distribution of the Wald from the N bootstrap samples.

We note that the auxiliary model used is a VAR(1) and is for a limited number of key variables: the major macro quantities which include GDP, consumption, investment, inflation and interest rates. By raising the lag order of the VAR and increasing the number of variables, the stringency of the overall test of the model is increased. If we find that the structural model is already rejected by a VAR(1), we do not proceed to a more stringent test based on a higher order VAR[13].

Rather than focus our tests on just the parameters of the auxiliary model or the impulse response functions, we also attach importance to the ability to match data variances, hence their inclusion in α. As highlighted in the debates over the 'Great Moderation' and the recent banking crisis, there is a major concern over the scale of real and nominal volatility. In this way our test procedure is within the traditions of RBC analysis.

We refer to the Wald statistic based on the full set of variables as the Full Wald test; it checks whether the a vector lies within the DSGE model's implied joint distribution and is a test of the DSGE model's specification in a wide sense. We show where in the Wald bootstrap distribution the Wald based on the data lies (the Wald percentile). We 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.[14]

We also consider a second Wald test, which we refer to as a 'Directed Wald statistic'. This focuses on more limited features of the structural model. Here we seek to know how well a particular variable or limited set of variables is modelled and we use the corresponding auxiliary equations for these variables in the VAR as the basis of our test. For example, we may wish to know how well the model can reproduce the behaviour of US output and inflation by creating a Wald statistic based on the VAR equation for these two variables alone.

References

Basawa IV, Mallik AK, McCormick WP, Reeves JH, Taylor RL (1991) Bootstrapping unstable first-order autoregressive processes. Ann Stat 19:1098–1101

[13] This increasing stringency is illustrated by the worsening performance of the model tested in Table 7 below for higher order VARs, as noted in footnote 7.

In fact the general representation of a stationary loglinearised DSGE model is a VARMA, which would imply that the true VAR should be of infinite order, at least if any DSGE model is the true model. However, for the same reason that we have not raised the VAR order above one, we have also not added any MA element. As DSGE models do better in meeting the challenge this could be considered.

[14] 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.

Bernanke BS, Gertler M, Gilchrist S (1999) The financial accelerator in a business cycle framework, Handbook of Macroeconomics. In: Taylor JB, Woodford M (eds), vol 1, ch. 21. Elsevier, pp 1341–1393

Canova F (1994) Statistical inference in calibrated models. J Appl Econ 9:S123–144

Canova F (1995) Sensitivity analysis and model evaluation in dynamic stochastic general equilibrium models. Int Econ Rev 36:477–501

Canova F (2005) Methods for Applied Macroeconomic Research. Princeton University Press, Princeton

Christiano L (2007) Comment on ‘On the fit of new Keynesian models’. J Bus Econ Stat 25:143–151

Christiano LJ, Eichenbaum M, Evans CL (2005) Nominal rigidities and the dynamic effects of a shock to monetary policy. J Polit Econ 113(1):1–45

Clarida R, Gali J, Gertler ML (1999) The science of monetary policy: A new Keynesian perspective. J Econ Lit 37(4):1661–1707

Dai L, Minford P, 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

Dave C, Jong D, DN (2007) Structural macroeconomics. Princeton University Press

Del Negro M, Schorfheide F (2004) Priors from general equilibrium models for VARs. Int Econ Rev 45:643–673

Del Negro M, 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 Q2:21–37

Del Negro M, Schorfheide F, Smets F, Wouters R (2007a) On the fit of new Keynesian models. J Bus Econ Stat 25:123–143

Del Negro M, Schorfheide F, Smets F, Wouters R (2007b) Rejoinder to comments on ‘On the fit of new Keynesian models’. J Bus Econ Stat 25:159–162

Evans R, Honkapohja S (2005) Interview with Thomas J. Sargent. Macroecon Dyn 9:561–583. 2005

Faust J (2007) Comment on ‘On the fit of new Keynesian models. J Bus Econ Stat 25:154–156

Fernandez-Villaverde J, Rubio-Ramirez F, Sargent T, Watson M (2007) ABCs (and Ds) of understanding VARs. American Economic Review 97:1021–1026

Friedman M (1953) The methodology of positive economics, in essays in positive economics. University of Chicago Press, Chicago

Gallant AR (2007) Comment on ‘On the fit of new Keynesian models’. J Bus Econ Stat 25:151–152

Gourieroux C, Monfort A (1995) Simulation based econometric methods. CORE lectures series, Louvain-la-Neuve

Gourieroux C, Monfort A, Renault E (1993) Indirect inference. J Appl Econ 8:S85–S118

Gregory A, Smith G (1991) Calibration as testing: Inference in simulated macro models. J Bus Econ Stat 9:293–303

Gregory A, Smith G (1993) Calibration in macroeconomics. In: Maddala G (ed) Handbook of Statistics, vol 11. Elsevier, St. Louis, Mo., pp 703–719

Hansen BE (1999) The grid bootstrap and the autoregressive model. Rev Econ Stat 81:594–607

Hansen LP, Heckman JJ (1996) The empirical foundations of calibration. J Econ Perspect 10(1):87–104

Horowitz JL (2001a) The bootstrap. In: Heckman JJ, Leamer E (eds) Handbook of Econometrics, vol.5, ch. 52. Elsevier, pp 3159–3228

Horowitz JL (2001b) The bootstrap and hypothesis tests in econometrics. J Econom 100:37–40

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

Kilian L (2007) Comment on ‘On the fit of new Keynesian models’. J Bus Econ Stat 25:156–159

Le VPM, Meenagh D, Minford P, Wickens M (2010) Two orthogonal continents? testing a two-country DSGE model of the us and the eu using indirect inference. Open Econ Rev 21(1):23–44

Le VPM, Meenagh D, Minford P, Wickens M (2011) How much nominal rigidity is there in the US economy — testing a New Keynesian model using indirect inference. J Econ Dyn Control 35(12):2078–2104

Le VPM, Meenagh D, 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

Le VPM, Meenagh D, Minford P (2014) Monetarism rides again? US monetary policy in a world of Quantitative Easing, Cardiff working paper No E2014/22, Cardiff University, Cardiff Business School, Economics Section; also CEPR discussion paper 10250, CEPR, London

Le VPM, Minford P, Wickens M (2013) A Monte Carlo procedure for checking identification in DSGE models, working paper E2013/4, Cardiff Economics Working Papers, Cardiff University, Cardiff Business School, Economics Section; also CEPR discussion paper

Le VPM, Matthews K, Meenagh D, Minford P, Xiao Z (2014) Banking and the macroeconomy in China: A banking crisis deferred Open Econ Rev 25(1):123–161

Liu C, Minford P (2014a) Comparing behavioural and rational expectations for the US post-war economy. Econ Model 43(C):407–415

Liu C, Minford P (2014b) How important is the credit channel? An empirical study of the US banking crisis. J Bank Finance 41:119–134

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

McCallum BT (1976) Rational expectations and the natural rate hypothesis: some consistent estimates. Econometrica 44:4–52

Meenagh D, Minford P, Wickens MR (2009) Testing a DSGE model of the EU Using indirect inference. Open Econ Rev 20(4):435–471

Meenagh D, Minford P, Wickens MR, Xu Y (2015) Comparing Indirect inference and likelihood testing methods: asymptotic and small sample results. Working paper, Cardiff Economics Working Papers No E2015/4, Cardiff University, Cardiff Business School, Economics Section

Meenagh D, Minford P, Nowell E, Sofat P (2010) Can a real business cycle model without price and wage stickiness explain UK real exchange rate behaviour J Int Money Financ 29(6):1131–1150

Meenagh D, Minford P, Wickens MR (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

Minford P, Ou Z (2013) Taylor Rule or optimal timeless policy? Reconsidering the Fed's behavior since 1982. Econ Model 32(C):113–123

Minford P, Theodoridis K, Meenagh D (2009) Testing a model of the UK by the method of indirect inference. Open Econ Rev 20(2):265–291

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

Sims CA (1980) Macroeconomics and reality. Econometrica 48:1–48

Sims CA (2007) Comment on 'On the fit of new Keynesian models'. J Bus Econ Stat 25:152–154

Smets F, Wouters R (2003) An estimated dynamic stochastic general equilibrium model of the euro area. J Eur Econ Assoc 1(5):p1123–1175

Smets F, Wouters R (2007) Shocks and frictions in US business cycles: A Bayesian DSGE approach. Am Econ Rev 97:586–606

Smith A (1993) Estimating nonlinear time-series models using simulated vector autoregressions. J Appl Econ 8:S63–S84

Watson M (1993) Measures of fit for calibrated models. J Polit Econ 101:1011–1041

Wieland V, Wolters MH (2012) Forecasting and policy making. In: Elliott G, Timmerman A (eds) Handbook of economic forecasting, Vol. 2. Elsevier

Wickens MR (1982) The efficient estimation of econometric models with rational expectations. Rev Econ Stud 49:55–67

Wickens MR (2013) How useful are DSGE macroeconomic models for forecasting? Open Econ Rev 25:171–193