DISCUSSION PAPER SERIES

No. 7539

THE 'PUZZLES' METHODOLOGY: EN ROUTE TO INDIRECT INFERENCE?

Vo Phuong Mai Le, Patrick Minford and Michael R. Wickens

INTERNATIONAL MACROECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP7539.asp**

ISSN 0265-8003

THE 'PUZZLES' METHODOLOGY: EN ROUTE TO INDIRECT INFERENCE?

Vo Phuong Mai Le, University of Cardiff
Patrick Minford, University of Cardiff and CEPR
Michael R. Wickens, University of Cardiff, University of York and CEPR

Discussion Paper No. 7539
November 2009

Centre for Economic Policy Research
53–56 Gt Sutton St, London EC1V 0DG, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 7539

November 2009

ABSTRACT

The 'Puzzles' Methodology: en route to Indirect Inference?

We review the methods used in many papers to evaluate DSGE models by comparing their simulated moments with data moments. We compare these with the method of Indirect Inference to which they are closely related. We illustrate the comparison with contrasting assessments of a two-country model in two recent papers. We conclude that Indirect Inference is the proper end point of the puzzles methodology.

JEL Classification: C12, C32, C52 and E1
Keywords: anomaly, Bootstrap, DSGE, indirect inference, puzzle, US-EU model, VAR and Wald statistic

Vo Phuong Mai Le
Cardiff Business School
Cardiff University
Aberconway Building
Colum Drive
CARDIFF
CF1 3EU

Email: LeVP@cardiff.ac.uk

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=164365

Patrick Minford
Cardiff Business School
Cardiff University
Aberconway Building
Colum Drive
CARDIFF
CF1 3EU

Email: minfordp@cf.ac.uk

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=100320

Michael R Wickens
Cardiff Business School
Cardiff University
Aberconway Building
Colum Drive
CARDIFF
CF1 3EU

Email: mrw4@york.ac.uk

Submitted 03 November 2009

Paper written for Economic Modelling volume in honour of P.A.V.B Swamy. We thank Arnab Bhattarachee, Huw Dixon, George Evans, Paul Levine, David Meenagh, Jo Pearlman, Christian Thoenissen, Jiang Yue and participants at the CDMA Conference in St. Andrew's 2009, for helpful comments. This work was supported by the UK's Economic and Social Research Council under grants RES-165-25-0020 and PTA-026-27-1623.

A popular way to evaluate macro models is to ask whether they can 'solve puzzles'. A puzzle occurs when a model cannot replicate a 'fact' such as a correlation between two variables, like consumption and the real exchange rate — thus generally a reduced form relationship that is reliably supposed to be observed in various available samples of data.

To 'solve a puzzle' a model, when subjected to the shocks that under its theory are supposed to strike it, should generate under such stochastic simulation a relationship on average close to the one in the data, while also similarly generating other relationships close to those found in business cycle data. We call this procedure the 'puzzles methodology'.

This way of evaluating models has largely replaced that of assessing their econometric 'fit' to the data, such as measures based on likelihood. Instead, by comparing the model's simulated behaviour with regularities in the data, as in the example above, one is asking whether the model can replicate the mechanisms at work in the data as evidenced by these regularities. When one asks whether a model has a high likelihood, one is effectively asking whether it can make a good forecast of the current data conditional on lagged endogenous and current exogenous data; yet being able to forecast well in this way is by no means the same as being able to replicate, unconditionally, the behaviour of the data as captured by particular regularities. This is what the puzzles methodology aims to do; and it is an aim which we firmly support even if in what follows we will be criticising the manner in which it is done and suggesting alternative procedures.

The question we address in this paper is whether this procedure, as widely carried out at present, has any real scientific value, in the sense that it tests any null hypothesis in a rigorous statistical manner. The method is presented as if at least informally it constitutes a test of the model's ability to 'fit the key facts'. Thus some distribution of each fact may be shown, and notably its variance. This distribution is supposed to be produced under the null hypothesis being tested. The facts involved are suggested to be the key ones that a model must fit to be of 'use' or 'interest' in the context. Those for a closed economy could be correlations of variables such as inflation, interest rates and consumption with GDP at various lags, for example. For an open economy these might be supplemented with correlations between the real exchange rate and some of these variables. In sum we are presented with some facts in the data and some distribution, usually a variance, around the model's 'prediction' (i.e. average stochastic value) for each fact; then if the data facts lie inside these distributions, say at the 95% confidence level, it is suggested that the model fits.

We will ask many questions about this procedure. They fall into five main categories:

1) how does one select the 'shocks' relevant to the model?

2) how does one decide the scaling/distribution/time-series behaviour of the shocks?

3) how does one estimate the distributions of the key descriptors of the data — e.g. persistence, cross-correlations of variables — under the null?

4) how does one allow for jointness of these distributions (i.e. for the comovements of the descriptors' estimates)?

5) how does one decide whether a model, though rejected as a whole as mis-specified, has any ability to fit any particular key facts?

As we attempt to confront these questions we will find that the puzzles procedure fails to give convincing answers to them.

Let us begin with the answers that would typically be given by a puzzles-practitioner ('theorist' hereafter).

1) The shocks selected should be those the theorist believes exist. For example, many RBC theorists only consider the existence of productivity shocks. Others, for example using cash-in-advance models, may add money supply shocks; some others consider a government spending shock as a 'demand' shock.

2) The scaling (standard deviation) of the key shock can be chosen ad hoc to fit some basic variance, such as that of GDP. Particular other shocks can be scaled ad hoc to fit other variances. Time-series behaviour can be imposed ad hoc.

3) Distributions of the key descriptors — e.g. their standard deviation or the Impulse Response Functions (IRFs) of a shock — are obtained from some time-series representation (e.g. a VAR) of the data. (For IRFs some method of identifying the shock is used that is not model-specific.)

4) There is no consideration of the joint distributions of these descriptors — e.g. of the joint distribution of the persistence of output and inflation.

5) The model's simulated values for the chosen descriptors are compared with the data-generated values and their distributions. Provided the simulated values are within the 95% interval around the data-generated values, the model is to be accepted; if some lie outside then the others represent the 'aspects where the model succeeds'.

This account of the typical theorist's procedure is meant solely as a template for criticism; clearly not every theorist does all of these things. But our point is that all studies do one or more of these things, while many do most of them, and some do all of them. By construction if any does not do any of them, then it is not to be included in our criticisms.

Now let us consider whether these procedures make sense.

1) Shock selection Should the theorist be free to choose which shocks are to be included in the model? Take a DSGE model of say 5 endogenous variables and 5 equations. When confronted with the data this model's implied residuals can be estimated together with the residuals' time-series properties. These residuals are the 'shocks' it implies. Furthermore, unless the model includes all 5 of these shocks, 'stochastic singularity' occurs, whereby at least one variable is a deterministic combination of the others — necessarily this will be false unless the 'shock' in one equation is zero.

A way that has been suggested around this problem is to assume there are 'measurement errors' and that some of the shocks are accounted for by such errors. The meaning of a measurement error is strictly that a variable is mis-measured, not that it is different from the equation's prediction in true fact. However, the assertion that a variable such as consumption for example is mis-measured in such a way that had one used the model equation one would have obtained the correct measurement is incredible: the modeller would be saying, absurdly, that the DSGE model is the basis for measuring what is happening in the economy, regardless of what the statisticians collect from surveys etc.

In practice this is not being asserted. Instead the 'measurement error' treatment of shocks, under which they are effectively ignored in the stochastic simulation, amounts to assuming that the error should not be treated as a stochastic disturbance.

Of course since the 'data', however mis-measured, are the basis of these tests, usually we are forced to treat them as if they are reliable up to some order of magnitude within which we can work robustly. Nevertheless the data may be imperfectly measured in a way that cannot be avoided by this assumption. For example in the GDP market-clearing equation the elements when loglinearised might not add up exactly to GDP and the residual, which has to be a measurement error, could be regarded as an element that varied in a 'non-stochastic' way.

The question is: what meaning can be given to such a statement?

A reason for treating shocks as non-stochastic would be that they were 'one-off' shocks that would either never or very rarely be repeated — such as German reunification. But while this might be true of some episodes for certain shocks in certain periods, it is hard to argue that it is true for the whole sample for a particular shock to one of the model variables. A reasonable way to deal with such one-off events would be to control for them in both the data and the model. Thus one 'extracts' their effects from the data and from the model shocks, and then compares the model on the remaining shock components with the data excluding this one-off variable's impact.

In sum while some shocks may be one-off and non-stochastic, these should be dealt with by allowing explicitly for them in the data and model. This could include variables that are genuinely mis-measured such as the missing element in the GDP market-clearing example above. However, there must remain in any model a large scope for straightforward shocks, to be interpreted as omitted variables such as shocks to preferences and technology. A modeller must include these errors in the model if the model is to be taken seriously as an account of the data.

2) Shock size and behaviour: Now consider 2), the scaling (size of standard deviation) and distribution (shape and time-series features) of these shocks. It has become rather usual, and most explicit in Bayesian estimation, to impose these on the shocks, at least up to some range. The justification seems to be that as these shocks are part of the model, the modeller can make assumptions at will or perhaps in line with some other information. However, this strategy faces the problem that in general it will violate the shocks implied by the model and data, as just discussed. For any given sample of data and assumed model parameters (including shock parameters), the shocks can be backed out of the assumed model's solution and the data. Thus when confronting the data, the modeller has no real choice to make, once the model parameters are chosen.

Consider the paradox that would be created by assuming errors different from these 'true' errors when engaging in testing the model against the data in a particular sample. The modeller could by choosing errors at will find that there was a good match between the data moments and the model moments. Yet this could be due to the choice of errors. With the true errors no match might exist.

A particular case of this occurs with scaling. It is possible that the model greatly over- or under-predicts the data moments. Yet by scaling the errors this discrepancy can be removed. It might be argued

that this scaling is innocuous, because one is only interested in the cross-moments — e.g. correlations — that relate one variable to another. However, one cannot be sure what scaling of errors will do to the model relationships. In effect one is multiplying the true errors by a set of impact parameters that now enter the model in addition to existing parameters. Yet if so, these parameters also impact on other variables.[1]

3)–5) Choosing descriptors of the data and comparing them with model simulations: Plainly the choice of data descriptors must be limited in some way by relevance to the matter in hand — e.g. business cycle policy — since a model cannot explain everything; the more measures included the greater the certainty of rejection. Some measures of data variability and also of variable interconnection for a selection of key business cycle variables will be chosen — e.g. second moments and key IRFs. We have no problems in general with these choices.

Where we do have problems is with the treatment of the null which is the DSGE model being tested. Statistical procedures for testing a null hypothesis imply that the test distributions should be derived under the null. In some cases authors show distributions for the descriptors that are derived from the data — e.g. from the VAR. This may in certain circumstances generate a valid test statistic — e.g. when the VAR can under the null be assumed to have the same distribution as the DSGE model. However since the DSGE model simulations can be used to derive the DSGE model's own distributions there is in general no need to make this assumption.

A second problem concerns identification of IRFs of structural shocks. Here authors often (e.g. Christiano et al, 2007) use extraneous identifying restrictions that are arbitrary. An example is the assumption that a shock to interest rates has no contemporaneous (i.e. within the quarter) effects on prices or output; this enables one to identify a policy shock to interest rates as equal to a unit change in the VAR shock to interest rates (as the policy shock has no effects on any other macro variable). Of course this creates the problem that the Euler equations for consumption and investment should respond to lagged interest rates — inconsistently with the proper model structure.

Such arbitrary identification is unnecessary when the null is a DSGE model that identifies the structural shocks. The null should be used to identify IRFs when there is no data-generated identifying method — as there rarely is.

The final problem which we have found to be of great importance is that authors routinely neglect the joint distributions of their descriptors implied by the DSGE null. We can give a simple example in the case where the two descriptors are the persistence of inflation and interest rates. It would be common practice for authors to quote the data mean estimate of persistence and its standard deviation (let us say under the DSGE model as argued above it should be). Then if the model simulations generate a mean within each of these distributions at some, say 95%, confidence level the author would announce that the model 'fits the facts'. However, if we recall the Fisher equation, we will see that the persistence of inflation and interest rates will be highly correlated. Thus in samples created by the DSGE model from its shocks where inflation is persistent, so will interest rates be; and similarly when the former is non-persistent so will the latter tend to be. Thus the two estimates of persistence under the null have a joint distribution that reflects this high correlation. It is quite possible for each data-generated persistence parameter to lie with its own distribution taken alone while the two together lie outside this joint distribution.

This situation is illustrated below, where we show a three-dimensional figure for the parameter distribution of a VAR with just two parameters, for example inflation and interest rates regressed only on their own individual past (a diagonalised VAR). Suppose that the model distribution is centred around 0.5, and 0.5; and the data-based VAR produced values for their partial autocorrelations of 0.1 and 0.9 respectively for inflation and interest rates — the two VAR coefficients. Suppose too that the 95% range for each was $0 - 1.0$ (a standard deviation of 0.25) and thus each is accepted individually. If the parameters are uncorrelated across samples, then the situation is as illustrated in the diagram below.

[1] Let the model be given by $Ay_t = ME_t y_{t+1} + Ny_{t-1} + u_t$, where $u_t = \Phi u_{t-1} + \epsilon_t$. This can be transformed into $y_t = A^{-1}MB^{-1}y_t + A^{-1}NLy_t + A^{-1}u_t$; where L is the lag operator and B^{-1} is the forward operator leading the variable while keeping the date of expectations constant (here at t). Assume that the model satisfies the saddlepath Blanchard-Kahn conditions (with f forward and l backward roots), then we can rewrite it as $\prod_{i=1}^{f}(I - \gamma_i B^{-1}) \prod_{j=1}^{l}(I - \lambda_j L)y_t = K(L; M, N, A)u_t$. Here we note that K is a function of all the parameters of the model, as well as involving lags of the errors produced by the backward roots and current values of the errors produced by the forward roots. We can solve for y_t in terms of the current shocks and its own lagged values by projecting the forward roots onto the errors and then projecting all the backward roots, as well as error autoregressive roots, onto y_t. It is clear that the impact effect, just like the transmission effect, comes from the complete parameter set.

The height of the diagram shows the density of parameter combinations across the samples. Here the mean of each parameter's distribution remains constant regardless of the value of the other parameter. Of course the joint parameter combination will also be accepted because of this independence.

Now consider the case where there is a high positive covariance between the parameter estimates across samples. Thus suppose that in samples with high inflation autocorrelation we also find high interest rate autocorrelation (because of the Fisher effect perhaps). The lower figure below illustrates the case for a 0.9 cross-correlation between the two parameters. The effect of the high covariance is to create a 'ridge' out of the 'density mountain'. Hence at high values of the interest rate autocorrelation the mean of the inflation autocorrelation is now increased from 0.5; for example at an interest rate parameter of 0.9 the mean of the inflation parameter distribution will be 0.86; the distance of 0.1 from a mean of 0.86 is 3.04 standard deviations. Thus the joint parameter combination of $0.1, 0.9$ will be rejected even though individually the two parameters are accepted.

When there are numerous VAR parameters, each pair will have the characteristics just described, creating ridges in multiple dimensions. The joint distribution of the VAR parameters will clearly in general have many such ridges.

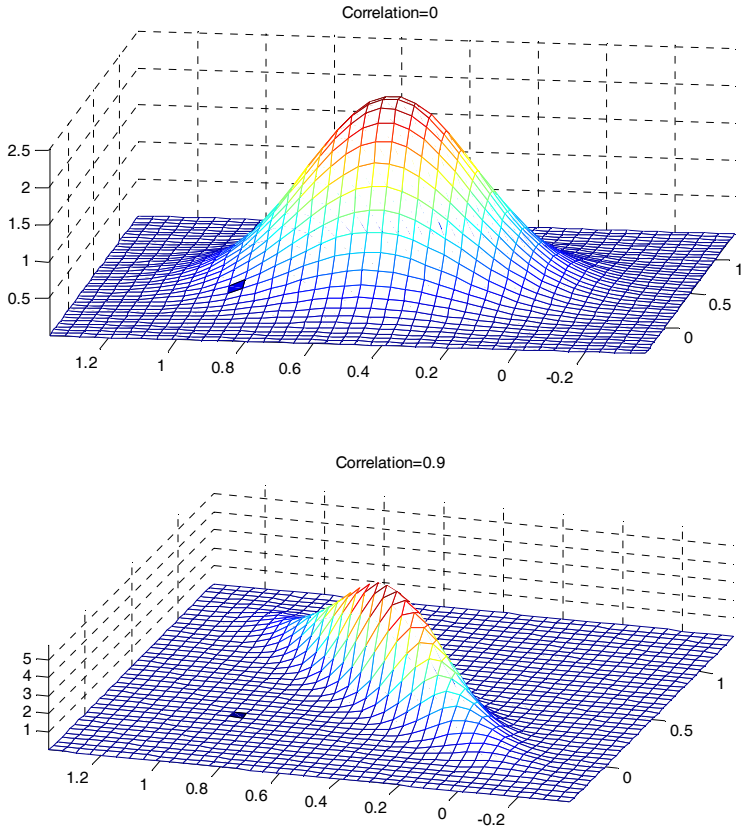

Figure 1: Bivariate Normal Distributions (0.1, 0.9 shaded) with correlation of 0 and 0.9.

The practical impact of this problem can be shown by considering that in Smets and Wouters' model of the EU using the errors implied by the data Meenagh et al (2008) found that if the VAR coefficients were used as descriptors then the model passed when these were considered individually but was rejected at 100% confidence when their cross-correlations were allowed for.

This problem remains when one is carrying out the evaluation of 'aspects where the model succeeds'. This evaluation too must take account of joint distributions of the chosen descriptors reflecting the various 'aspects'. The fact that a model fits say inflation IRFs individually by no means implies that it fits their joint distribution.

Conclusions on puzzles methods We conclude from this survey of popular methods in the puzzles literature that they do not necessarily establish anything with statistical confidence. This is not to say that they are useless; it may well be that exposing mechanisms in particular DSGE models and comparing them with broad facts of business cycle behaviour may lead to DSGE models that can be successfully tested against the data. However in the end we would suggest that they should be tested by methods based on statistical distributions generated by the null hypothesis.

1 An alternative method of pursuing the same objectives — Indirect Inference as a source of Strong Econometrics

We have proposed a method for model evaluation that had not been applied before: the method of indirect inference — Smith (1993), Gregory and Smith (1991, 1993), Gourieroux et al. (1993), Gourieroux and Montfort (1995) and Canova (2005). One may view the recent literature of Real Business Cycle (RBC) models as having a similar intention in its development of the Puzzles methodology we have just reviewed, in the sense that much of it simulates models and checks whether their selected unconditional moments and cross-moments are similar to those of the data.

The idea behind this comparison is to check whether a model can replicate dynamic features of the real world. This focus is to be contrasted with direct inference where models are compared with the data directly, using measures of fit related to the size and behaviour of residuals; thus Bayesian or Full Information Maximum Likelihood estimation in effect obtain the smallest possible set of current residuals. Models that fit well may nevertheless fail to behave dynamically like the data; and vice versa, as is fairly obvious; fitting data conditional on lagged endogenous and current exogenous variable values is not the same as replicating patterns of behaviour over time unconditionally. The former relates to the predictive efficiency of the model. The latter asks whether the model could be the generating mechanism of the data — the most fundamental test of all.

This dichotomy in model performance was not so clear with 'first generation' macro models whose theoretical underpinnings were loosely specified (the models in Bryant et al, 1993, are examples). These models were, by virtue of this looseness, able to build in current and lagged relationships that could both achieve good fits to the data and generate plausible dynamic behaviour. However there has been general dissatisfaction with the ad hoc nature of their assumptions as a guide to understanding the effects of policy and indeed of particular large shocks (such as the banking crisis) that could also cause shifts in behaviour — a view first set out in Lucas' critique of macroeconometric models (Lucas, 1976). This dissatisfaction led to the current dominance in macroeconomic analysis of 'second generation' DSGE models with tight micro-foundations that could withstand this critique.

However for these models the empirical dichotomy is pronounced. These models are generally unable to fit the data closely because of their tight specifications; and the question that has been asked of them is whether they can nevertheless fit the dynamic features of the data. We can think of a DSGE model as capturing some essential economic mechanisms driving the effects of policy and other shocks; plus some set of independent shocks which can be driven by their own exogenous processes.

From this point we can examine the two main alternative approaches to testing these models mentioned in the introduction: the 'Puzzles' and the 'Strong Econometrics' methodologies. We have reviewed the Puzzles methodology above and identified serious difficulties with it from an econometric viewpoint.

The Strong Econometrics approach (the name was coined by Smets and Wouters, henceforth SW; examples are SW, 2003 and 2007; and Christiano et al, 2005) by contrast as its name implies attempts to apply formal statistical techniques to the data and models. The pioneering work of SW involved Bayesian methods, under which the model parameters and error properties are formulated as prior distributions, with posterior values estimated that take account of the data. Following standard econometric procedures every relationship is given a shock considered due to omitted variables; these variables are much like the errors in standard econometrics, in that they represent factors affecting variables' behaviour that could occur within the model's theoretical framework but cannot be modelled. Thus they could be 'interpreted' (for example as shocks to preferences, current and future expected technology) but are exogenous and cannot be directly observed or modelled, except as residuals. Thus the model is expected to generate stochastically separate outcomes for all observed endogenous variables, which is necessary if the model is to have any chance of formally fitting the facts about these variables. (This avoidance of 'stochastic singularity' distinguishes this treatment from the Puzzles methodology where endogenous variables not subject to a shock are exact linear combinations of the other endogenous variables that are subject to shocks.)

It should be noted that SW's work does not, and is not intended to, test the models. It assumes the structure to be basically correct; the matters of interest are which values within the range of the priors get closest to the data, thus telling us the best version of the model and whether it is of a particular nature (e.g. has more or less price/wage rigidity). One could thus characterize SW's Bayesian methods, with apologies to Mary Poppins, as 'supercalibration' in the sense that they aim to improve on straight calibration from priors by taking account of the data within fairly tight ranges around the calibrated priors. Our proposal by contrast has been to test the models, in this respect more like the Puzzles methodology, with a view to rejecting some and, we would hope, accepting others. To do this with statistical rigour we propose the existing theoretical econometric approach of Indirect Inference. Under this approach one formulates an auxiliary model whose role is to describe the data in a manner that is neutral between the competing structural, here DSGE, models in contention. For example this might well be a Vector Auto Regression (VAR). Then one simulates the behaviour of the DSGE model and obtains the implied auxiliary model that emerges from these simulations. Under Indirect Inference estimation one may alter the DSGE parameters to get the simulated and actual auxiliary models as close as possible. Under Indirect Inference evaluation one may simply check whether the two models are statistically close enough to avoid rejection at some specified confidence level.

Our testing procedure under Indirect Inference treats the strongly estimated model as the null hypothesis, the true description of the economy. By implication the residuals in the structural equations are the true errors, and therefore the source of the sampling variation for the model. Any particular sample episode will under this null be a solution of the model with a drawing from these errors. We exploit this in our procedure, using the bootstrap to obtain the small sample distributions implied by the model. Notice that by implication we should not use any assumed errors in our test because under the null, these will not be the true errors. Of course if the modeller has prior information about these errors, then comparison of this with the true model errors will already constitute a test of the model; however, such information on errors is rare, it would seem — unlike perhaps parameters, where micro studies can give fairly clear priors in some cases.

It might seem that to assume any DSGE model is the true model is absurd, given the heavy abstraction involved in their specification. A number of authors have used measures of 'closeness' to the data for models that they assume are false — see for example recently Bhattacharjee and Thoenissen (2007) who use a measure based on the var-covar matrix of the data, and Canova and Sala (2009) for a critique of the closeness measure proposed by Del Negro et al (2006). We should be clear that we follow the classical econometric testing procedure which assumes that the model is true in order to create the test, since without a null hypothesis no testing statistic is possible; thus we ask whether the model can be rejected or not.[2] The Puzzles methodology essentially has to make the same assumption; the only difference is that it tests the DSGE model by informal rather than formal statistical means.

We evaluate the models as a whole using a Wald statistic which is derived from the joint distribution generated by the DSGE model for the chosen features of the data. This joint distribution is calculated from the model bootstraps (i.e. sample replications created by drawing from the structural errors) very simply; each bootstrap yields a set of the chosen features (obtained by estimating these features on that sample) so that the joint distribution of these is given by the values obtained from all the bootstraps. The Wald statistic is calculated as the percentile band around the joint mean within which the data-generated features fall in that joint distribution. Thus a number such as 96 would indicate that these fall beyond the 95% 'contour' of the joint distribution, and the model would then be rejected by the data at the 95% confidence level.

We also develop a set of 'directed' Wald statistics where we evaluate the model on more restricted sets of features, using exactly the same principle. This can guide us to which features of the data the model can capture, which not; and therefore indicate what sort of respecification would help or alternatively which conditional predictions of the model can be safely used for policy purposes.

In order to carry out the Indirect Inference testing procedure we need to make a variety of practical choices. First, we need to decide what features of the data are to be considered of major importance for the test; plainly, the wider the net is cast, the more likely a model is to fail. For this reason we restrict ourselves to a subset of major variables — such as output, inflation, interest rates, consumption, and the exchange rate — and we consider first their variances as an indispensable measure of the model's capacity to capture variability, and second their VAR coefficients as a measure of the model's capacity

[2] Unless DSGE models are to be considered untestable, this procedure is inevitable. When authors assert that their model has large and implausible residuals on certain equations which they propose to ignore in testing it, they are effectively saying that their model is econometrically untestable because they are refusing to set up an explicit null hypothesis from their model.

to capture dynamic interactions between them. We also look at traditional measures used in model evaluations, such as cross-moments and impulse response coefficients; but we note that these are derived from the VAR coefficients and so can be considered redundant under our procedure.

Second, current Indirect Inference procedures have been worked out for asymptotic properties under stationarity. In further work we are considering how the procedures can be extended to non-stationary data. However, for all our work here we have worked with stationarised data. To do this we have reported results with a variety of filtering methods.

1.1 Indirect Inference as Strong Econometrics illustrated and compared with Puzzles

In this section we briefly describe results we obtained for a two-country model of the US and the EU using data from 1975–2000. We then go on to compare these with the results reported by Chari Kehoe and McGrattan (2002) for a two-country model for essentially the same period where these authors use the puzzles methodology described above. We note a number of discrepancies between their and our findings; and we conclude by suggesting some possible reasons.

1.1.1 Models of the EU and US singly and as a two-country world

Our work began with a DSGE model created by SW (2003, 2007) along lines similar to that of Christiano et al. (2005) This model can be considered to embody most of the key features that New Keynesian (NK) authors have found useful in fitting the facts — backward-looking indexation in sticky wage and price equations, habit persistence in consumption, variable capacity utilization and a q-theory of investment. SW estimated this model by Bayesian methods for both the EU and the US, treating each as a closed economy. We took these models as represented by their mean posterior parameters and subjected them to testing by Indirect Inference in a series of recent papers — Meenagh et al (2008) 'EU paper', Le et al (2008) 'US paper' and Le et al (2009) 'EU-US paper'. For this purpose we used the data of SW for the EU, which were filtered by the least intrusive available method, linear detrending. It turns out that provided one uses a different trend for each series this is adequate to create stationarity for virtually all these series. We found the same was true for US data and used the same filter.

We also considered as a robustness check filtering the EU data by the Hodrick-Prescott filter. This gave very similar results — annexes I and J of Supporting Annex for the EU paper. We further checked on EU data whether any difference would come from filtering out via dummy variables of episodes with major shocks such as German reunification, accession of new EU countries and oil shocks. Annexes K and L reveal that again these dummies made no difference to the test results.

To assist in focusing the test of these models we created a variant benchmark New Classical (NC) version of the same model with minimal nominal rigidity provided by a one quarter information lag in household labour supply. We also varied the model's interest rate (Taylor) rule, choosing parameters that performed better (in the Indirect Inference sense) with the NC version. SW did not include such parameter values in their prior distributions, yet the assumption of NK rigidity with backward indexation seemed to us to be a matter of controversy, especially given the results above on UK inflation persistence.

We first carried out the procedures described above on the EU and US models separately. We consider their performance in two main dimensions: their capacity to replicate a) the variability and b) the VAR coefficients of the data. What was striking was that neither the NK nor the NC model could remotely replicate the data variability. NK models produce too little variability in nominal variables (inflation and interest rates) and too much in real variables (output, consumption and investment). NC models produce the opposite — too much for nominal, too little for real.

We therefore posited a set-up in which each household sold its labour (and similarly each intermediate firm sold its output) to two markets, imperfectly and perfectly competitive, in a fixed proportion — this weight was potentially different in the labour and product market. We also assumed that the Taylor Rule would be a weighted combination of the two ones appropriate to NK and NC. We searched for the best weights in the Indirect Inference manner, seeking those that would generate a fit to the data variability. It turned out that quite low NK weights (in the range of 5–10% in both labour and product markets in the EU, and 10–20% in the US) were sufficient to get the weighted model to fit this well. Small weights generate a disproportionate degree of inflation smoothing because their indirect effect via expected inflation multiplies the smoothing power.

We also investigated a two-country version of these models, where we joined together the EU and US models by allowing households in each to buy the goods from the other and hold the nominal

bonds (not equity or other assets) of the other. This gives rise to two new first-order conditions: one giving substitution between home and foreign goods, the second giving the uncovered interest parity equation. Because each country pursues differing Taylor Rules reacting to its own shocks, the exchange rate insulates each country's households from the other's bond market where different interest rates hold; risk sharing is therefore greatly limited.[3]

We explored the capacity of the weighted models both alone and in this two-country version to fit the VAR coefficients. Here we found a striking feature of our tests, that we had not expected. The models could fit the vast majority of VAR coefficients individually with fair comfort which led us to expect them to fit jointly. However, all versions of the model were totally rejected on the Wald test on the joint distributions.

We then tried to locate the source of this rejection by calculating 'directed' Wald statistics where we narrowed the focus to selected variables in the two-country model. We found that when only output and the real exchange rate were considered, the model was accepted; however, when nominal variables were considered, it was rejected. This suggested to us that the source of the model's problems lay in the assumption of a simple Taylor rule unvarying over the whole period which for the two-country model was from 1975 to 1999. It seems likely that around the mid-1980s there was a shift towards less accommodative monetary policy in both the EU and the US, following the period in the early 1980s when central banks and governments focused their efforts on bringing 'double-digit' inflation down into low single digits. Previous work on the UK (Minford et al, 2009) where careful modelling of monetary regime shifts seemed to pay dividends reinforced this idea that careful modelling of monetary policy and its various shifts could much improve the EU-US model's nominal performance.

1.1.2 Some details of the Indirect Inference results for the two-country model

We now review in some more detail the results we obtained for the two-country model described above. We start with the model's overall fit.

It turns out that the combined weighted model performs reasonably well also as one would expect from the models' individual success. The model's predicted variance bounds embrace the data variances. While as with both the individual economy models the Full Wald statistic rejects at 100, the model's Average t-test Wald statistic (including data variances) is 82.7, indicating substantial closeness to the data — remembering that this is average closeness of each feature taken individually, much as is done is the current literature comparing models with data moments etc. The components in this are shown in the Table below; at its foot are the data variances and their 95% bounds.

[3] We can examine the effects of degrees of market completeness as follows, following CKM (2002).

Let us suppose in a two-country world (say the US and the EU) that there are open capital markets across borders- that is there are no exchange controls so that people in country A can buy the assets of country B. As noted by Chari, Kehoe and McGrattan a complete contingent asset available in just one of the countries then can be bought by people in both countries and by arbitrage it follows that the real exchange rate (ξ) between the two countries equals the ratio of the marginal utilities of consumption (U_c) in every state (s):

$\xi_t(s_t) = U^*_{ct}/U_{ct}(s_t)$

If there are only non-contingent bonds available then the equivalent condition (ignoring second order terms) is that the expected change in the log of the real exchange rate equals the log change in the ratio of the marginal utilities of consumption:

$E_t \ln \xi_{t+1} - \ln \xi_t = E_t \ln(U^*_{ct+1}/U_{ct+1}) - \ln(U^*_{ct}/U_{ct})$

A last point we can make is that whether there are fully complete markets via complete contingent loans or merely non-contingent loans with incomplete markets makes essentially no formal difference to the model: thus under complete markets we have $\ln \xi_t = \ln u_t$ (where u_t is the ratio of consumption marginal utilities) while under incomplete we have $E_t \ln \xi_{t+1} - \ln \xi_t = E_t \ln u_{t+1} - \ln u_t$ or $(1 - B^{-1}) \ln \xi_t = (1 - B^{-1}) \ln u_t$ where B^{-1} is the forward operator under expectations at t; and thus again $\ln \xi_t = \ln u_t$ (+ constant of integration) since the forward operator expressions cancel. The addition of the constant is of no consequence for the evaluation of responses to shocks as used here. Intuitively we can say that the only way the two variables can follow exactly the same path in response to a shock from one period to the next is if they are equal at all points.

Notice that whether the degree of risk-sharing between the two countries' consumers depends on how far the real exchange rate is stabilised by the model including policy reactions.

	Actual	Lower	Upper	State
$A^{Y^{US}}_{Y^{US}}$	0.848156	0.734829	1.154908	IN
$A^{\pi^{US}}_{Y^{US}}$	0.007809	−0.02349	0.189169	IN
$A^{R^{US}}_{Y^{US}}$	0.022051	−0.03326	0.127268	IN
$A^{NE}_{Y^{US}}$	0.012121	−0.82331	1.022503	IN
$A^{Y^{EU}}_{Y^{US}}$	−0.01314	−0.19684	0.179005	IN
$A^{\pi^{EU}}_{Y^{US}}$	0.006146	−0.12742	0.102156	IN
$A^{R^{EU}}_{Y^{US}}$	−0.01151	−0.14166	0.069589	IN
$A^{RXR}_{Y^{US}}$	0.038521	−0.3344	0.587617	IN
$A^{Y^{US}}_{\pi^{US}}$	0.36066	−0.77054	0.109068	OUT
$A^{\pi^{US}}_{\pi^{US}}$	0.52233	0.414022	0.830549	IN
$A^{R^{US}}_{Y^{\pi US}}$	0.291259	0.002895	0.312914	IN
$A^{NE}_{\pi^{US}}$	−1.53716	−0.61581	3.010255	OUT
$A^{Y^{EU}}_{\pi^{US}}$	−0.07586	−0.35787	0.421549	IN
$A^{\pi^{EU}}_{\pi^{US}}$	0.376145	−0.23157	0.178844	OUT
$A^{R^{EU}}_{\pi^{US}}$	0.063915	−0.19015	0.166679	IN
$A^{RXR}_{\pi^{US}}$	0.82309	−0.25858	1.4899	IN
$A^{Y^{US}}_{R^{US}}$	−0.36486	−1.40835	0.284688	IN
$A^{\pi^{US}}_{R^{US}}$	0.159178	−0.82883	0.053359	OUT
$A^{R^{US}}_{R^{US}}$	0.799615	0.015509	0.702261	OUT
$A^{NE}_{R^{US}}$	0.859896	−5.91636	1.220997	IN
$A^{Y^{EU}}_{R^{US}}$	0.075664	−0.87739	0.617746	IN
$A^{\pi^{EU}}_{R^{US}}$	−0.00657	−0.56982	0.337031	IN
$A^{R^{EU}}_{R^{US}}$	0.01617	−0.45617	0.371871	IN
$A^{RXR}_{R^{US}}$	$-4.20E-05$	−3.33276	0.11577	IN
$A^{Y^{US}}_{NE^{US}}$	−0.014	−0.19792	0.24851	IN
$A^{\pi^{US}}_{NE}$	0.002316	−0.07708	0.155404	IN
$A^{R^{US}}_{NE}$	−0.00233	−0.08023	0.093029	IN
A^{NE}_{NE}	0.91884	−0.04945	1.899053	IN
$A^{Y^{EU}}_{NE}$	−0.0045	−0.22315	0.17572	IN
$A^{\pi^{EU}}_{NE}$	0.001812	−0.174	0.09997	IN
$A^{R^{EU}}_{NE}$	−0.00059	−0.16908	0.066184	IN
A^{RXR}_{NE}	0.043954	−0.38027	0.567782	IN
$A^{Y^{US}}_{Y^{EU}}$	0.136033	−0.3435	0.312359	IN
$A^{\pi^{US}}_{Y^{EU}}$	0.016644	−0.27198	0.058838	IN
$A^{R^{US}}_{Y^{EU}}$	−0.03228	−0.18539	0.045186	IN
$A^{NE}_{Y^{EU}}$	0.747852	−1.57235	1.10588	IN
$A^{Y^{EU}}_{Y^{EU}}$	0.981749	0.522227	1.067224	IN
$A^{\pi^{EU}}_{Y^{EU}}$	0.046921	−0.23584	0.116394	IN
$A^{R^{EU}}_{Y^{EU}}$	0.04625	−0.20801	0.105403	IN
$A^{RXR}_{Y^{EU}}$	−0.55428	−0.75171	0.520743	IN

Table 1: VAR coefficients and variances for the SW EU-US weighted model

	Actual	Lower	Upper	State
$A^{Y^{US}}_{\pi^{EU}}$	−0.19749	−1.46488	0.66343	IN
$A^{\pi^{US}}_{\pi^{EU}}$	0.234097	−1.36936	−0.20396	OUT
$A^{R^{US}}_{\pi^{EU}}$	0.00023	−1.0349	−0.19345	OUT
$A^{NE}_{\pi^{EU}}$	0.262419	−8.32418	1.148038	IN
$A^{Y^{EU}}_{\pi^{EU}}$	0.19797	−0.16902	1.761419	IN
$A^{\pi^{EU}}_{\pi^{EU}}$	0.363756	0.10701	1.286825	IN
$A^{R^{EU}}_{\pi^{EU}}$	−0.05942	−0.23094	0.745215	IN
$A^{RXR}_{\pi^{EU}}$	−0.8632	−4.77715	−0.22719	IN
$A^{Y^{US}}_{R^{EU}}$	−0.23979	−0.65119	1.62167	IN
$A^{\pi^{US}}_{R^{EU}}$	−0.19154	0.300122	1.516317	OUT
$A^{R^{US}}_{R^{EU}}$	0.158096	0.226208	1.149287	OUT
$A^{NE}_{R^{EU}}$	−2.64786	−0.56199	9.468748	OUT
$A^{Y^{EU}}_{R^{EU}}$	−0.50621	−2.01799	0.122446	IN
$A^{\pi^{EU}}_{R^{EU}}$	0.034226	−0.7663	0.407617	IN
$A^{R^{EU}}_{R^{EU}}$	0.860364	−0.23734	0.784417	OUT
$A^{RXR}_{R^{EU}}$	0.326105	0.710531	5.647189	OUT
$A^{Y^{US}}_{RXR}$	0.032076	−0.47781	0.446239	IN
$A^{\pi^{US}}_{RXR}$	−0.01252	−0.23182	0.235984	IN
$A^{R^{US}}_{RXR}$	−0.00368	−0.11584	0.241804	IN
A^{NE}_{RXR}	−0.29155	−1.84215	2.255991	IN
$A^{Y^{EU}}_{RXR}$	−0.01729	−0.36227	0.451164	IN
$A^{\pi^{EU}}_{RXR}$	0.017608	−0.17549	0.363608	IN
$A^{R^{EU}}_{RXR}$	−0.00042	−0.1415	0.338608	IN
A^{RXR}_{RXR}	0.920969	−0.1275	1.867483	IN
$\sigma^2_{Y^{US}}$	5.945734	3.493806	36.88094	IN
$\sigma^2_{\pi^{US}}$	0.385286	0.301302	0.788716	IN
$\sigma^2_{R^{US}}$	0.806059	0.230181	0.824717	IN
σ^2_{NE}	351.4456	48.61269	530.9087	IN
$\sigma^2_{Y^{EU}}$	3.608468	1.49129	11.46202	IN
$\sigma^2_{\pi^{EU}}$	0.245893	0.221595	0.855852	IN
$\sigma^2_{R^{EU}}$	0.363569	0.197476	0.751455	IN
σ^2_{RXR}	43.54279	9.932144	95.91408	IN

Table 2: VAR coefficients and variances for the SW EU-US weighted model (cont.)

What is striking is how little spill-over there is between the US and EU economies. A variance decomposition shows that while both economies' shocks affect the real exchange rate and the trade balance, only the home shocks affect home macro variables, much in the way they did when each works as a closed economy. We can also see that both models give a preponderant effect to productivity and labour supply shocks, with no demand shocks contributing much of the variation in either inflation or output in the US. Only in the EU does the monetary shock contribute an important share of output and inflation variation.

Shock↓\Variable⟶	Y^{US}	π^{US}	R^{US}	NE	Y^{EU}	π^{EU}	R^{EU}	RXR
$Prod^{EU}$	0.003	0.015	0.024	10.445	23.895	7.397	19.670	10.078
$Cons^{EU}$	0.000	0.000	0.000	0.006	3.057	8.975	7.906	0.234
Res^{EU}	0.000	0.000	0.001	0.162	2.533	0.446	0.916	0.151
Inv^{EU}	0.000	0.003	0.005	1.479	5.512	2.500	6.706	1.489
Mon^{EU}	0.003	0.011	0.018	6.894	16.732	61.357	26.844	6.774
$Price^{EU}$	0.000	0.000	0.000	0.050	0.037	4.962	2.481	0.055
$LabSup^{EU}$	0.009	0.029	0.047	20.891	48.233	14.361	35.470	19.844
$Wage^{EU}$	0.000	0.000	0.000	0.000	0.000	0.002	0.004	0.000
Res^{US}	1.031	1.535	2.809	1.328	0.000	0.000	0.000	1.547
$Cons^{US}$	0.612	2.229	2.695	0.145	0.000	0.000	0.000	0.658
Inv^{US}	2.441	1.878	3.994	1.871	0.000	0.000	0.000	1.921
Mon^{US}	0.310	5.877	0.416	0.594	0.000	0.000	0.000	0.612
$Prod^{US}$	31.489	28.209	29.697	17.531	0.000	0.000	0.001	19.357
$Price^{US}$	0.792	1.336	0.726	0.334	0.000	0.000	0.000	0.311
$Wage^{US}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$LabSup^{US}$	63.308	58.876	59.566	38.271	0.001	0.001	0.002	36.967
$Total$	100	100	100	100	100	100	100	100

Table 3: Variance Decomposition for US-EU weighted Model

2 Evaluating dimensions of the model's performance

As we have just seen from the variance decomposition this model says that the domestic performance of the EU and US economies is entirely determined by domestic shocks in each case and that these shocks are predominantly supply shocks. Only in the EU is there an important role for a monetary demand shock. All these shocks do however impact on the real exchange rate (and trade balance) of course, as the 'buffer' between the two economies.

This can allow us to evaluate the model's ability to replicate certain aspects of the data. First of all, we can limit the number of variables we examine and estimate a VAR for that group of variables alone, and compute the Wald statistic for its VAR coefficients. Then we can also ask whether the model's IRFs for a particular shock are rejected by the data. Thus we evaluate the IRFs for say the productivity shock on output and inflation by grouping the average IRFs for the US productivity shock on US output and inflation, and those for the EU productivity shock on EU output and inflation; these make 4 average IRFs. We then find the joint distribution of these average IRFs to compute the Wald statistic for the data-observed average IRFs.

We call these 'Directed Wald' statistics. The idea is that they give us a guide to where the model is mis-specified — or, looked at another way, to what uses in analysis we can safely put the model.

What we find — Table 4 below — is that the model passes these Wald tests at the 95% level for outputs and the real exchange rate, but fails them when interest rates or inflation are added. This pinpoints the DSGE model's failure in respect of nominal relationships (echoing the fact that of the VAR coefficients lying outside their 95% bounds, many involve inflation and interest rate effects). Thus the model, which indeed has properties essentially like a real business cycle model, can replicate the behaviour of real variables well but fails on nominal variables. This suggests rather clearly that monetary relationships in the model require attention — perhaps not surprisingly given the turbulence in monetary policy during this era.

If we turn to shocks — Table 5 — we find that US and EU productivity and labour supply shocks all pass their individual Directed Wald tests quite comfortably when interest rate effects are excluded; but when interest rates are included only productivity shocks do. Thus we can see here again that while the shocks do well for most variables they fail on interest rates. The US monetary shock only contributes non-trivially to the variance of US inflation, whose IRF to it comes comfortably within its 95% bounds. The EU monetary shock explains a big part of the variation in EU inflation, whose IRF to it lies within its 95% bounds. What all this shows is that taken individually the shock effects are well-modelled on the whole, apart mostly from their effects on interest rates; but as we know from the tests on variables, taken together they cannot account for nominal behaviour. Apart from failing on interest rates, they fail when combined even when interest rates are excluded, because of the general failure to pick up inflation effects.

Finally, we note that the model fits the data variances as a group at the 95% level (Table 6), provided

NE is excluded. Although the model can match the NE variance singly, it cannot match it jointly, no doubt because it is hugely variable. We do not attach much importance to this variable as it is the difference between the natural logs of US exports to the EU and EU exports to the US; this trade is actually quite small, specialised, and volatile. We can omit it from the VAR and the results are essentially unchanged.

Summarising these tests of particular capacities of the model, we can say that it fits the data variability, it is capable of replicating output and real exchange rate behaviour, but that it fails on nominal variables. If one prefers to evaluate a model through its shock IRFs then we can say the model captures the effects of the important shocks individually in both economies on output and inflation, but not generally on interest rates; furthermore in an echo of the results for VAR coefficents when all the important shocks in the model are examined together for their effects on all major variables, with or without interest rates, their IRFs are jointly rejected. Thus this model can be rigorously tested econometrically for certain properties and it passes some of these limited tests and gives us some clues about the location of mis-specifications.

Variable combinations	Direct Wald
Y^{EU}, Y^{US}	80.7
π^{EU}, π^{US}	99.9
R^{EU}, R^{US}	96.1
Y^{EU}, Y^{US}, RXR	94.2
$Y^{EU}, Y^{US}, RXR, R^{EU}, R^{US}$	100
$Y^{EU}, Y^{US}, RXR, R^{EU}, R^{US}, \pi^{EU}, \pi^{US}$	100
RXR $(AR(2))$	81.3

Table 4: Directed Wald Statistics by variable combinations

Shocks	Variables	Direct Wald
Mon^{EU}	Y^{EU} R^{EU} π^{EU} RXR	84.1
$Prod^{EU}$	Y^{EU} π^{EU} RXR $(R^{EU} incl)$	47.1 (86.5)
$LabSup^{EU}$	Y^{EU} π^{EU} RXR $(R^{EU} incl)$	66.1 (96.1)
$Prod^{US}$	Y^{US} π^{US} RXR $(R^{US} incl)$	77.5 (87.7)
$LabSup^{US}$	Y^{US} π^{US} RXR $(R^{US} incl)$	94.9 (99.7)
$Prod^{BOTH}$	*without (with) interest rates*	64.1 (97.3)
$LabSup^{BOTH}$	*without (with) interest rates*	93.0 (100)
$(Prod, LabSup)^{BOTH}$	*without (with) interest rates*	100 (100)

Table 5: Directed Wald Statistics by shocks

Variances of data				Direct Wald
$\sigma^2_{Y^{EU}}$	$\sigma^2_{Y^{US}}$	$\sigma^2_{\pi^{EU}}$	$\sigma^2_{\pi^{US}}$	94.3
$\sigma^2_{R^{EU}}$	$\sigma^2_{R^{US}}$	σ^2_{RXR}		

Table 6: Directed Wald Statistic for Variances of the Data

3 What does this model tell us about the world economy?

3.1 The nature of the world economy

This model suggests the world is bi-polar: the US and the EU are essentially independent blocs with little mutual spill-over. This can be illustrated by deterministic productivity shocks for each bloc shown below.

This bi-polarity implies that the exchange rate acts as a buffer between the two blocs enabling each to achieve its own market-clearing real interest rates under its own policy preferences. Hence its wild swings as uncorrelated shocks come from each direction. This can be seen in the IRFs below for the exchange rate reaction. Notice that a positive supply shock in one economy causes its real interest rates to fall to clear the goods market and so its real exchange rate to depreciate; thus a US productivity rise

causes the US real exchange rate to depreciate (RXR to fall). This is what we see for example in the data-based IRFs (coming from the data VAR as indentified by the model) for a US productivity shock. Thus the data too can be seen through the lens of this model as supporting this interpretation.

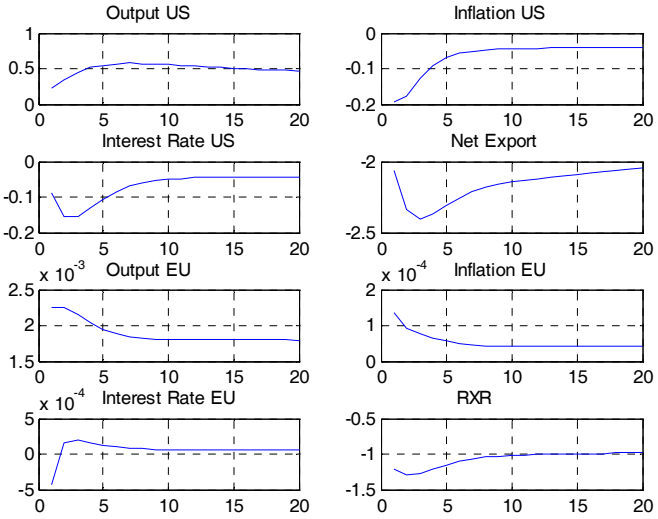

Figure 2: US productivity shock

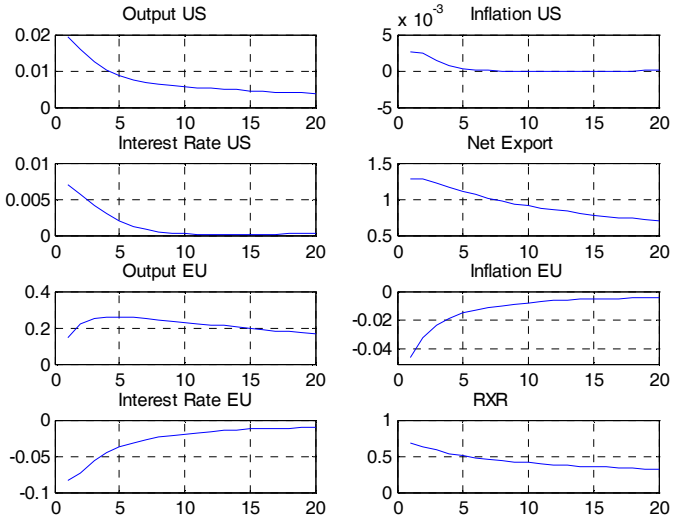

Figure 3: EU productivity shock

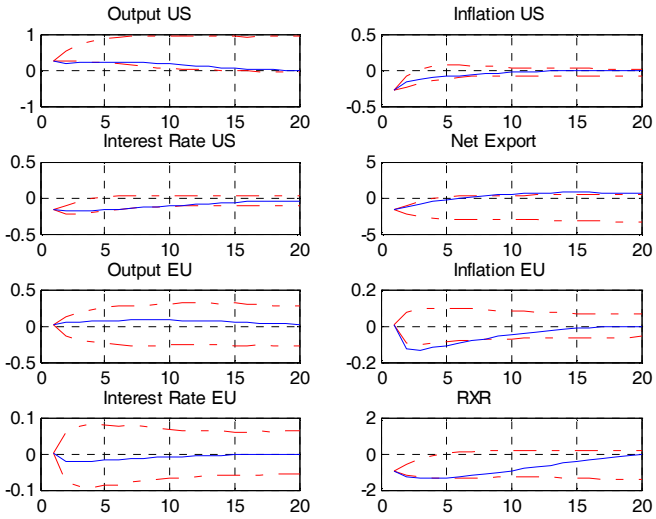

Figure 4: VAR IRFs for US productivity shock with model bounds

3.2 Contrasting results from Puzzles papers

We now proceed to review papers on a two-country model for the EU and the US over essentially the same period. We focus in particular on Chari, Kehoe and McGrattan (2002), CKM, as a well-crafted example; however the remarks we make about this paper are typical of those we would make about other ones using the same methods and they appear to apply equally to papers that take the same approach as CKM to the same data — see also Henriksen, Kydland and Sustek (2008), and Kollmann (2009) for recent work in this area.

In their paper CKM calibrate a model of the US and Europe in which wage and price rigidity, adjustment costs in investment, habit persistence in consumption, and incomplete asset markets are all assumed at various stages. This makes their model similar to the SW model considered above — the main difference being the absence of variable capacity utilisation which on its own seems unlikely to be a major difference. CKM's aim is primarily to replicate the autocorrelation and volatility of prices, the nominal and the real exchange rate in the data but they are also concerned to replicate some business cycle moments — variability of consumption/investment/net exports relative to GDP, their autocorrelations and their cross-correlations. Their conclusion is that they can match the business cycle behaviour reasonably but are left with a minor anomaly that the model cannot quite generate enough pesistence in the real exchange rate and a key anomaly, that the real exchange rate is highly positively correlated with relative consumption in the model but hardly at all in the data.

Of course it is impossible to examine their results using our methods without full stochastic simulation as above. However, their model is as we have seen in essence the same as the original (NK) version of SW, for which we have a full set of results. These we can compare with those reported by CKM using puzzles methods.

Let us take in turn CKM's findings: a) that the model matches business cycle behaviour, b) that it just fails to match the real exchange rate's persistence c) that it fails massively to match the absence of correlation between the real exchange rate and relative consumption.

a) We find that the SW model we have used to represent CKM's broadly replicates their results on business cycle variables. Thus the Table below finds as they do in Table 6 of their paper that the SW model matches a set of business cycle moments. For all but two of the moments shown the model mean bootstrap estimate is within two (model-generated) standard deviations of the data values. This is actually rather better than CKM find for their model, so if anything it strengthens their claim that models of this general type do well in matching the business cycle individual measures.

	Data	Model bootstrap Estimate	(Model bootstrap Stdev)
Standard deviations Relative to GDP			
US			
Consumption	1.1324	1.0411	(0.1414)
Investment	4.1514	2.4789*	(0.4564)
Net exports	3.3390	3.7587	(1.1972)
EU			
Consumption	1.2313	1.0692	(0.0881)
Investment	3.2080	3.9733	(0.5338)
Autocorrelations			
US			
GDP	0.9362	0.9331	(0.0340)
Consumption	0.9484	0.9400	(0.0345)
Investment	0.9697	0.9379	(0.0290)
Net exports	0.9356	0.9132	(0.0460)
EU			
GDP	0.9502	0.8724	(0.0598)
Consumption	0.9651	0.8643	(0.0624)
Investment	0.9638	0.9269	(0.0353)
Cross-Correlations			
between foreign and domestic			
GDP	0.4449	0.1282	(0.3551)
Consumption	0.0261	0.0096	(0.3749)
Investment	−0.2342	0.0611	(0.3339)
between net exports and US GDP	−0.1043	−0.4916	(0.2842)
between real exchange rate and			
US GDP	−0.2312	−0.3687	(0.3185)
Net exports	−0.1653	0.9786*	(0.0177)
* outside 2-standard-deviation bounds		Wald statistic	100

Table 7: Business cycle statistics for the model

Yet in contrast to CKM's claims, while the SW model of the US and the EU matches these business cycle moments individually as shown, it does not does not match the *joint* behaviour of real business cycle variables at all well; notice that the Wald statistic for the joint distribution of all these descriptors is 100, indicating total joint rejection.

Looking at the model's performance in more detail, we first look at whether the model can match the data variances. As can be seen below in the table of variances and model bounds, it is hopelessly unable to match the nominal variables with the single exception of US inflation; for EU inflation and both countries' interest rates the model produces much too low a variability.

Secondly, we use the VAR as a summary of the variables' comovements and ask whether the joint distribution of the VAR coefficients according to the SW model embraces the data-generated VAR coefficients within the 95% boundary. The resulting Wald should be less than 95. When all variables, nominal and real, are included in the VAR, the Wald is 100. The same goes for the VAR coefficients involving inflation and interest rates; many of these are rejected individually — the rejected coefficients are listed below and as can be seen all of them involve inflation or interest rates.

	Y^{US}	π^{US}	R^{US}	NE	Y^{EU}	π^{EU}	R^{EU}	RXR
actual	5.9457	0.3853	0.8061	66.2868	3.6085	0.2459	0.3636	43.5428
lower	3.7058	0.1609	0.1118	33.4875	1.1035	0.0346	0.0437	7.3960
upper	32.5493	0.4157	0.3635	352.0475	6.6040	0.0732	0.1043	64.9710
mean	12.9236	0.2716	0.2138	122.7267	3.0342	0.0522	0.0691	24.5612

Table 8: Variances

	Actual	Lower	Upper	State
$y^{US} - \pi^{US}$	0.36066	−1.00144	−0.16673	OUT
$NE^{US} - \pi^{US}$	−0.66758	0.532847	3.66168	OUT
$\pi^{EU} - \pi^{US}$	0.376145	−0.10909	0.105305	OUT
$R^{EU} - \pi^{US}$	0.063915	−0.09618	0.063754	OUT
$R^{US} - R^{US}$	0.799615	0.22052	0.68045	OUT
$R^{EU} - y^{EU}$	0.04625	−0.04711	0.037255	OUT
$\pi^{EU} - \pi^{EU}$	0.363756	−0.34765	0.072481	OUT
$\pi^{US} - R^{EU}$	−0.19154	−0.17685	0.546092	OUT
$y^{EU} - R^{EU}$	−0.50621	−2.48253	−0.95338	OUT
$R^{EU} - R^{EU}$	0.860364	0.435151	0.749359	OUT

Table 9: VAR coefficients for US-EU weighted Model

Notice also in passing how important the 'mark-up' shocks on wages and prices are for both the US and the EU — see variance decomposition below. As pointed out in CKM (2008) it is hard to give these shocks a serious interpretation. Fortunately as we found above, when one moves to a model version that has far less nominal rigidity these shocks become largely irrelevant.

Shock↓\Variable⟶	Y^{US}	π^{US}	R^{US}	NE	Y^{EU}	π^{EU}	R^{EU}	RXR
$Prod^{EU}$	0.012	0.044	0.078	26.064	49.719	0.471	44.269	28.515
$Cons^{EU}$	0.000	0.000	0.004	0.126	5.855	0.018	7.372	0.099
Res^{EU}	0.000	0.000	0.001	0.208	4.370	0.007	0.613	0.228
Inv^{EU}	0.003	0.011	0.017	3.988	14.356	0.171	10.119	4.396
Mon^{EU}	0.008	0.009	0.040	12.749	25.220	0.677	35.572	14.082
$Price^{EU}$	0.000	0.000	0.001	0.241	0.454	97.321	1.537	0.272
$Wage^{EU}$	0.000	0.000	0.001	0.036	0.024	1.333	0.512	0.040
Res^{US}	1.781	0.422	2.927	1.227	0.000	0.000	0.000	1.872
$Cons^{US}$	1.043	0.435	3.480	0.020	0.000	0.000	0.000	0.368
Inv^{US}	7.237	2.777	8.573	0.936	0.000	0.000	0.000	1.453
Mon^{US}	1.784	4.038	18.108	2.832	0.000	0.000	0.000	2.960
$Prod^{US}$	54.469	3.854	27.261	33.758	0.001	0.000	0.005	30.996
$Price^{US}$	25.922	63.907	29.222	12.153	0.000	0.000	0.000	10.562
$Wage^{US}$	7.739	24.503	10.288	5.662	0.000	0.000	0.000	4.156
$Total$	100	100	100	100	100	100	100	100

Table 10: Variance Decomposition for US-EU weighted Model

Eliminating nominal variables and leaving only real GDP of the US and the EU plus the real exchange rate, we obtain an acceptance with a Wald of 90 — thus we can say that real business cycle behaviour in the broadest measure is matched by the model. But when we add consumption and investment into this VAR, the Wald rises again to 100. The model cannot match the joint cyclical behaviour of consumption, investment and GDP.

Variables	Directed Walds
Y^{EU}, Y^{US}, RXR	90.2
$Y^{EU}, Y^{US}, RXR, C^{EU}, C^{US}, INV^{EU}, INV^{US}$	100

Table 11: Directed Walds

Why is this so? It seems that a variety of things come into play. First, we use shocks estimated directly from the data and the model. Second, we use all the shocks in the model's behavioural equations. Third, we look at joint distributions of behavioural coefficients and not single distributions of each individually; as we have seen the joint distributions are far more demanding than the totality of the single distributions. Of these three reasons it seems that the last is the main one, since we managed to generate — similarly to CKM — matching of the individual moments using all the shocks in the model and directly estimated from the data/model.

When we come to b) which is a single distribution, we do however find that the model easily matches the persistence of the real exchange rate, taken on its own — again when we apply the method of Indirect Inference we find that the result contradicts CKM. The reason in this case appears to be a failure to generate the model's bounds using the shocks coming from the data — no joint distributions is at issue here.

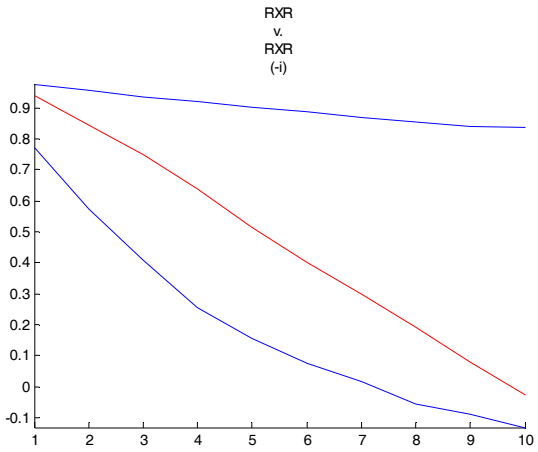

Figure 5: RXR auto-correlation

Unfortunately this success of the model does not occur jointly with success in matching the business cycle. There are many models that can match the real exchange rate's persistence on its own. For example it is matched as easily by the weighted SW model we describe above and this model does a better job of matching the business cycle data, particularly data variances.

Finally, we consider c), CKM's 'key anomaly'. Here they are on strong ground. The SW NK model generates as one would expect powerful positive correlations between the real exchange rate and relative consumption. The lower 95% bound of these does indeed lie well above the data cross-correlation.

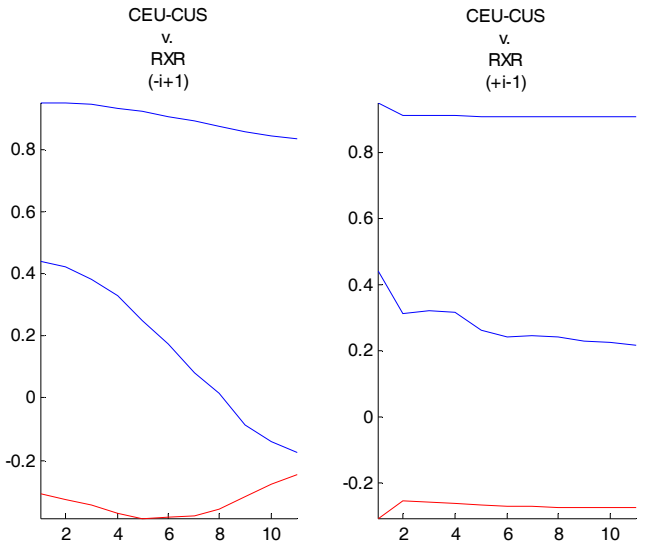

Figure 6: Correlation of $C^{EU} - C^{US}$ with RXR

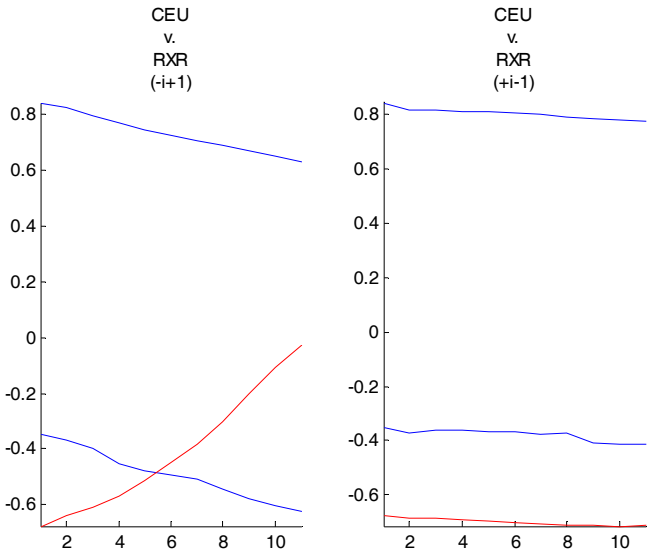

Figure 7: Correlation of CEU with RXR

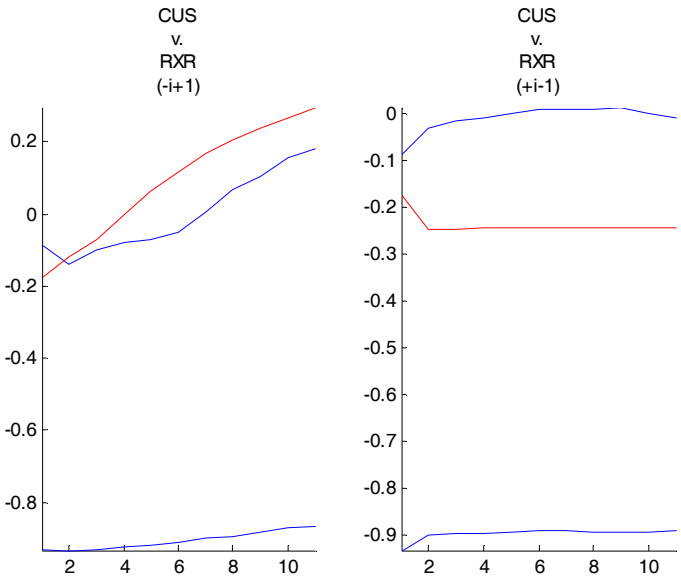

Figure 8: Correlation of CUS with RXR

CKM suggest in response to these results that the model's problem in matching the consumption-RXR behaviour is to be contrasted with its ability to match consumption and investment's business cycle behaviour. However we found above that the model actually fails to match this business cycle behaviour; there is therefore no such contrast. The model's failure to match the consumption-RXR behaviour is of a piece with its failure to match consumption and investment behaviour generally. Also the directed Wald of the two consumptions and the RXR is 99.6, confirming that the model fails to capture interrelations of these variables. Mainly the problem arises in the EU; RXR and CUS relations are fairly well captured by the US model as shown by the directed Wald (93.6) and the chart below. As RXR rises CUS falls, as it should. However, the problem lies with RXR and CEU; as RXR rises CEU falls even more, contrary to the model theory; the directed Wald is 98.5.

Cons/RXR	Directed Walds
RXR, C^{EU}, C^{US}	99.6
RXR, C^{EU}	98.5
RXR, C^{US}	93.6

Table 12: Directed Walds

The problem this model has with matching the correlation between RXR and relative consumption is quite a general problem for these models. Thus it applies to the best weighted version we examined above in essentially the same way. Thus these models do have trouble matching the detailed interactions of business cycle variables, basically because they impose a host of tight restrictions and to do so generate errors in the model first order conditions which are interpreted as the effects of omitted variables. If within these errors are concealed further interactions that the model is not picking up, then this mis-specification will produce lack of dynamic matching, as we see here. Also when mis-match is so general, it seems hard to justify emphasising any particular mis-match such as the RXR correlation with relative consumption. In short, the problem highlighted by CKM's 'key anomaly' is the same as the one we have highlighted in its failure to match the joint behaviour of business cycle variables: the model imposes filigree restrictions on joint behaviour of economic variables which simply cannot be found in the data.

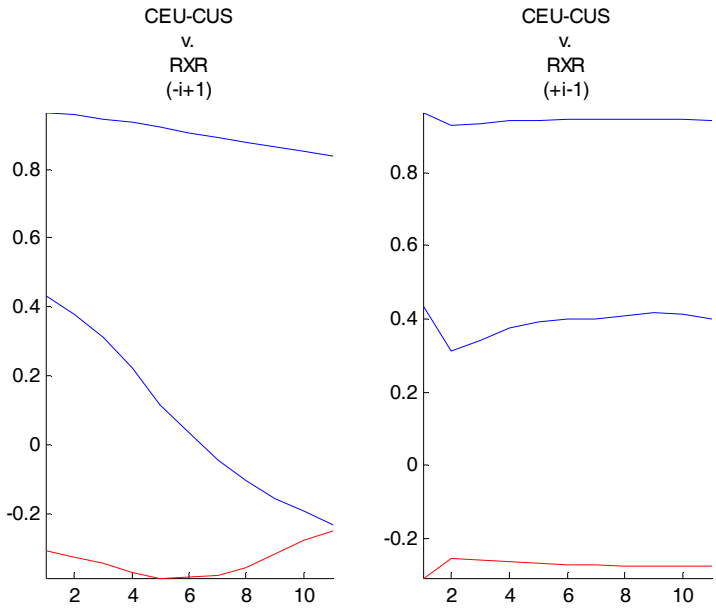

Figure 9: Correlation RXR with $C^{EU} - C^{US}$ (Weighted Model)

4 Conclusions

Summarising, we can say that CKM's type of model with nominal wage/price rigidity, as represented here by SW's NK model, is rejected by the data on both the cyclical behaviour of consumption and on the comovements of the real exchange rate and relative consumption. The model is in fact able to match the cyclical behaviour of output and the real exchange rate, including the latter's persistence. Nevertheless, the model is quite unable to match the behaviour of nominal variables and so it seems unlikely that nominal rigidity via its effects on nominal variables can be the reason for exchange rate behaviour.

Our main aim here however has not been so much to highlight these substantive model assessments but rather to illustrate the pitfalls of the puzzles methodology in assessing model performance. We have argued that the model's shocks should be estimated from the data and not imposed; and that model properties should be assessed against the data using their joint distributions which generally pose more stringent requirements than their single distributions viewed collectively. In the one case where CKM look at a joint property — viz the correlation of the real exchange rate with the relative country consumptions, a compound of its correlation with each individually — they found strong rejection. Joint distributions reflect the filigree restrictions imposed by DSGE models on relationships between variables; such restrictions are not easily met in the data.

We end by noting with approval the basic aim of puzzles method authors which is to evaluate a DSGE model against the data by comparing the model's simulated behaviour with data descriptors rather than by the 'goodness of fit' on residuals used in 'direct' estimation and inference. In this approach puzzles authors in effect are using a sort of indirect inference. The plea of this paper is that the method should be used formally as illustrated here. We conclude that the puzzles methodology is a method en route to indirect inference: the sooner it arrives the better for our understanding and assessment of macro models.

References

[1] Bhattacharjee, A. and C. Thoenissen (2007) 'Money and monetary policy in dynamic stochastic general equilibrium models', Manchester School, 75 (s1), 88-122.

[2] Canova, Fabio. 2005. *Methods for Applied Macroeconomic Research.* Princeton: Princeton University Press.

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

[4] Chari, V., P. J. Kehoe and E. McGrattan (2002) 'Can sticky price models generate volatile and persistent real exchange rates?' Staff Report 277 from Federal Reserve Bank of Minneapolis, published in Review of Economic Studies, Vol. 69, No. 3, July 2002, pp. 533-563.

[5] Chari, V., P. J. Kehoe and E. McGrattan (2008) 'New Keynesian Models: Not Yet Useful for Policy Analysis' 14313, NBER Working Paper 14313, National Bureau of Economic Research, Inc. published in American Economic Journal: Macroeconomics,American Economic Association, vol. 1(1), pages 242-66, January.

[6] Christiano, Lawrence, Martin Eichenbaum and Charles L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," Journal of Political Economy, University of Chicago Press, vol. 113(1), pages 1-45, February.

[7] Del Negro,M., F. Schorfheide, F.Smets, R. Wouters (2006) 'On the fit of new Keynesian models', Journal of Business and Economic Statistics, 25,143–162.

[8] Gregory, Allan W. and Gregor W. Smith. 1991. "Calibration as testing: Inference in simulated macro models." *Journal of Business and Economic Statistics*, 9: 293–303.

[9] Gregory, Allan W. and Gregor W. Smith. 1993. "Calibration in macroeconomics" In *Handbook of Statistics*, ed. G. Maddla,11: 703-719. St. Louis, MO.: Elsevier.

[10] Gourieroux, Christian and Alain Monfort. 1995. "Simulation Based Econometric Methods." CORE Lectures Series, Louvain-la-Neuve.

[11] Gourieroux, Christian, Alain Monfort and Eric Renault. 1993. "Indirect inference." *Journal of Applied Econometrics*, 8: 85–118.

[12] Henriksen, E., Finn E. Kydland, and Roman Sustek (2008) 'The High Cross-Country Correlations of Prices and Interest Rates', mimeo, Bank of England

[13] Kollmann, Robert (2009) 'Household Heterogeneity and the Real Exchange Rate: Still a Puzzle', CEPR Discussion Paper No 7301, C.E.P.R., London.

[14] Le, M., P. Minford and M. Wickens (2008) "How much nominal rigidity is there in the US economy? Testing a New Keynesian DSGE Model using indirect inference" Cardiff University Economics Working Paper: http://www.cf.ac.uk/carbs/econ/workingpapers/papers/E2008_32.pdf.

[15] Le, M., D. Meenagh, P. Minford and M. Wickens (2009) "Two Orthogonal Continents: Testing a Two-country DSGE Model of the US and EU Using Indirect Inference" Cardiff University Economics Working Paper: http://www.cf.ac.uk/carbs/econ/workingpapers/papers/E2009_3.pdf.

[16] Meenagh, D., P. Minford and M. Wickens (2008) "Testing a DSGE model of the EU using indirect inference", forthcoming Open Economies Review (available online, DOI: 10.1007/s11079-009-9107-y), Cardiff University Economics Working Paper: http://www.cf.ac.uk/carbs/econ/workingpapers/papers/E2008_11.pdf

[17] Minford, P., K. Theodoridis, and D. Meenagh (2009) "Testing a model of the UK by the method of indirect inference" Open Economies Review (2009) Vol. 20(2), April, pp. 265-291; available as Cardiff University Economics Working Paper at http://www.cf.ac.uk/carbs/econ/workingpapers/papers/E2007_2.pdf

[18] Smith, Anthony. 1993. " Estimating nonlinear time-series models using simulated vector autoregressions." *Journal of Applied Econometrics*, 8: S63–S84.

[19] Smets, F. & R. Wouters, 2003. "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area," Journal of the European Economic Association, MIT Press, vol. 1(5), pages 1123-1175, 09.

[20] Smets, F. & R. Wouters, 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," American Economic Review, American Economic Association, vol. 97(3), pages 586-606, June