1 Introduction: the transmission mechanism and monetary policy

Lavan Mahadeva and Peter Sinclair

The transmission mechanism of monetary policy explains how monetary policy works – which variables respond to interest rate changes, when, why, how, how much and how predictably. This broadens to the issue of what monetary policy can do and what it should do to offset the effects of disturbances on inflation.

This volume sets out how the transmission mechanism is analysed for the purpose of informing monetary policy. The chapters that follow tackle different aspects of how a central bank can build a good working understanding of the transmission mechanism of monetary policy. In this introduction, we summarise how this understanding relates to the forecast apparatus and models employed, along with practical difficulties to be overcome. We highlight two key aspects of the monetary transmission mechanism – the monetary sector and the exchange rate – and conclude by summarising the key elements of current good practice.

1.1 How does the central bank analyse the transmission mechanism?

A central bank's interest in the transmission mechanism of monetary policy arises from the fact that it takes time for monetary policy to exert its maximum impact on inflation.[1] A central bank has to know how to position its interest rate *now* to keep inflation *in the future* close to its target, while avoiding any excessive destabilisation of output. It also has to form some view about what might happen to inflation and output over this intervening period (see Blinder, 1998; Budd, 1998).

If resources permitted, the central bank would be continually constructing and revising a comprehensive quantitative picture of the transmission mechanism of monetary policy over the policy horizon of, say, one or two years. Ideally, the monetary policy forecast would encompass much more than just predicted outcomes for inflation and output. It would include the following elements:

1. A set of *models* of the transmission mechanism with an explanation of how each model can be consistent with the others.

2. A set of *values for the underlying processes* that drive the forecast. These values are not just the exogenous variables but also the parameters and the off-model adjustments to endogenous variables. In the 1999 version of the Bank of England's macroeconometric model, for example, inventories as a ratio of GDP are determined by an exogenous de-stocking trend that partly accounts for innovations in integrated production processes and stock management techniques (Bank of England, 1999b, p. 50). In principle, this trend can be altered in the forecast, for example to reflect a view that the technical change has run its course, leaving the inventory–output ratio at a new constant equilibrium. The coefficient on any variable in the equation can also be thought of as a driving process; there may be good evidence to suggest that the firm's behaviour has altered, shifting the elasticities from their past values (Bank of England, 1995, p. 24).[2] Inventories could also be affected by other factors outside those in the equation. These factors could still be allowed to influence the inventory forecast as off-model adjustments.
3. Assumptions about *the policy reaction function*. What economic data do the central bank's policy decisions respond to, and how? What are the public's expectations of that policy reaction function?
4. By combining 1, 2 and 3, *the predicted out-turns* for inflation and output as well as for any other endogenous variable that can be matched against data when they become available.
5. Since forecasts can be wrong, *an appreciation of why previous forecasts turned out wrong* and what that could mean for the future forecast.

Inflation-targeting central banks, and many others too, make their views about what is likely to happen over this horizon explicit.[3] The reason is that credibility can be acquired and retained through the provision of transparent explanations, by publishing an inflation report or a forecast, for example (Mishkin and Posen, 1997; Bernanke et al., 1999; Chortareas et al., 2001).

The information presented becomes important if the central bank's view over this horizon is formed to *explain* and not just to *inform* policy. The public can become party to the uncertainties associated with the transmission mechanism. Higher moments, such as the risks and variances, would be quantified and published along with the expected forecast values. The presentation becomes an account of how the forecast depends on the interpretation of the current set of data. The weight attached to individual data series in determining forecasted inflation and output is not constant; it varies from policy round to policy round as new information comes to light (Svensson, 1999).

How is this understanding of the transmission mechanism over the future horizon to be built? For the purposes of illustration, we can split the construction of this forecast into the following stages:[4]

1. Formulate various hypotheses about how the driving processes (exogenous variables, residual adjustments and parameters) are developing and will develop.

2. Test the hypotheses against each other, comparing how they match up to theory and data. This testing can be done in two stages. Each hypothesis would first be cast into a theoretically consistent model. Then the predictions of this model would be compared with recent data. The most favoured hypotheses provide the expected values (and also the associated probability distribution of outcomes) of the driving processes in the forecast.
3. Calculate the implications for inflation and output using a model. These implications are conditional on assumptions about both the policy reaction and the expectations of this policy reaction.
4. Derive the optimal policy response that would return inflation to target without incurring unsatisfactory output losses.[5]

The key in specifying the driving processes – step 1 – is establishing how they might deviate in the future from their past behaviour. Shocks to the driving processes ultimately originate in shifts to the parameters that shape the preferences of agents, the micro-structure of markets, technological developments or endowments. The categories of agents whose preferences can shift are investors, borrowers, consumers, workers, firms and the monetary and fiscal authorities. The micro-structure of markets where goods, services, financial and non-financial assets, factors of production and information are traded can also change. Technological developments can arise in the production of goods, services, financial and non-financial assets, working capital (physical and human), inputs and information. Finally changes can occur in the endowments of natural resources, technical progress and the size of the working population, all of which are typically taken to be exogenous but not, of course, necessarily constant over time. A shock can also arise from changes in population composition, because different cohorts or groups may vary in preferences and constraints.

Once we have located an underlying cause, we still have to determine other dimensions of each shock. To begin with, the time path of the underlying disturbance needs specification. Is it permanent or temporary? If temporary, does it die off gradually or sharply? As shocks affect different sectors of the economy in different ways, does a sectoral or disaggregated model tell us significantly more about monetary policy implications than an aggregate model?[6] For open economies, the international dimension of shocks can be crucial. The implications of a disturbance can differ greatly, depending on whether it occurs only at home or is mirrored in the country's trading partners or capital partners abroad.

Can the shock be satisfactorily understood in isolation or is it related to other concurrent developments? Even if two disturbances have the same underlying cause, this relationship need not be an explicit part of the model used to forecast inflation. Consider an IT improvement that is skill biased, not only raising the overall level of productivity but also shifting demand from unskilled to skilled labour. We can capture how this skill-biased IT shock hits the economy in a model that disaggregates skilled and unskilled labour markets.

But, in more aggregate models, we could also combine an 'adjustment' to the wage equation with an exogenous shift to the time profile of technical progress.

No single methodology or model can convincingly analyse all shocks in this degree of detail. One solution is to build our picture of the transmission mechanism from individual studies covering different aspects of individual shocks. As an example, Peersman and Smets' contribution to this volume (chapter 2) explores how a common monetary policy tightening affects the euro area and how it depends on the economic cycle. Their findings are an important step in quantifying the money transmission mechanism of the region as a whole.

By introducing a common monetary shock into the models for growth of eight euro-area countries,[7] Peersman and Smets estimate that monetary policy exerts a significantly stronger effect when the euro area is in recession compared with a boom. Although the asymmetry varies among the countries sampled, on average it seems to be only two-fifths as strong in an expansion as in a recession. The authors explain that asymmetric monetary policy transmission could come about because economic conditions influence the sensitivity of consumption and investment decisions to monetary policy shocks through the extension of credit to constrained borrowers and investors. An alternative explanation is that nominal prices were stickier in expansionary than in disinflationary episodes (Ball and Mankiw, 1994). Other research by Peersman and Smets (Peersman and Smets, 2000b) isolates and tests the credit-availability effect on data disaggregated by sector, which could be more revealing. These findings are relevant for euro-area monetary policy. More generally, they remind us that shocks are not always linear in effect. Linearisation is a convenient assumption since the effects of each shock will then be independent of other events. Each shock can be judged in isolation and the effects of different shocks can then be summed to give us an aggregate picture of the transmission mechanism. Although linearity may often be a sufficiently close approximation,[8] the policy implications derived from linear models can sometimes be misleading. Furthermore, the interaction of a set of shocks makes their combined impact differ from that resulting from adding their effects one by one.

In this book, as elsewhere, a wide variety of models and approaches are now used to analyse shocks in such detail. But the tools at our disposal, and the models on which they are based, are consistent with the following two tenets:

- The economy converges towards a long-run state in which prices are fully flexible in nominal terms. The long-run state is described by an adapted neoclassical model in which the adaptations allow for rigidities in *real* prices that prevent the markets for goods and factors of production from fully clearing.[9]
- In the short run, the nominal prices of goods, inputs and assets are not fully flexible. New Keynesian phenomena, which arise chiefly because of menu costs, imply some degree of predetermination in nominal prices.[10] Any shock

to a nominal variable (for example, a surprise monetary policy action) will not be accompanied by an instantaneous complete adjustment to all other nominal variables. Relative prices, and hence real quantities, will not always be at the values they would take if all nominal prices were flexible – nominal shocks have temporary real effects. The scale of the real effects arising from nominal rigidities is in part determined by the real rigidities (Ball and Romer, 1991). So, even if monetary policy-makers cannot affect these underlying long-run market failures, the need to research them can still exist.

Models of the transmission mechanism that are grounded on this consensus can be labelled as 'IS–LM plus Phillips curve' descriptions of the economy. As the name implies, any model of this class can be summarised into a simple aggregate structure comprising only an IS curve, a policy rule and a Phillips curve[11] after many simplifying assumptions[12] are made.

1.2 How can we derive the policy implications?

To give policy advice, the driving processes, and the 'IS–LM plus Phillips curve' models in which they are couched, are combined. One way of combining is to ask whether the underlying shock implies any movements in the *real rate of interest*, assuming that the nominal interest rates and the expectations of it are constant.[13] A large enough forecasted fall (rise) in the real interest rate, when the nominal rate is unchanged, would imply that inflation will shift above (below) its target. If so, nominal interest rates would have to change.

No matter how this is done, when a response is called for, the path for setting interest rates depends on the temporary output losses and the preferences of the central bank concerning these costs. The output losses arise from both the original shock and the policy response to it. A major enterprise of monetary policy then becomes to identify the implications for the real interest rate and for real output of any shock, and to ensure that these estimates are correctly conditional on unchanged policy.

A popular scheme for deriving the policy implications of shocks has been to classify them into demand- and supply-side shocks, on the basis of what they imply for real output. Given constant expectations of monetary policy, inflationary demand-side shocks are associated with a rise in temporary output that leaves long-run or potential output unaffected, whereas inflationary supply-side shocks are associated with a fall in potential output with actual output falling or constant. Monetary policy should react to the shocks that are identified as inflationary on the demand side, whereas it is likely to respond to supply-side shocks only if the inflationary threat outweighs the output costs of responding. If the output implications of a shock are visible before its inflationary impact, this 'demand- versus supply-side' scheme can, in principle, serve as an early warning system for monetary policy.[14]

To put this into practice, the monetary policy-maker must be able to separate the output consequences of a shock into the changes in the potential value on the one hand and the changes in the output gap (actual output minus potential output) on the other. The relevant concept of potential output for monetary policy purposes here is the output that would arise at each moment in time if all nominal prices were flexible (McCallum and Nelson, 1999b; McCallum, 2000a,b). To apply this theory to the measurement of the current output gap, it is necessary to estimate what would determine output if prices were flexible: the counterfactual 'flexible-price' values of inputs into production, real rigidities (such as the factors that determine structural unemployment) as well as exogenous technical progress. Data that would determine the flexible-price value of inputs may be hard to come by. The flexible-price rate of unemployment, for example, is determined by structural factors such as workers' reservation wage rates and their bargaining power (see, for example, Layard, Nickell and Jackman, 1991, and Nickell, 1996). The data we have on the real unemployment benefit and union density can serve as a poor proxy for these deep structural factors. The importance of these theories has to be balanced with the implications of theories that link the flexible-price rate of unemployment to the actual rate of unemployment, such as hysteresis models (Cross, 1995; Ball, 1999a). Many researchers have relied instead on measures of potential output that employ only a few broad assumptions about the cyclical nature of potential output series compared with the output gap series. These measures tend to use only a single time series for output.[15]

As an alternative, Chadha and Nolan (chapter 3 in this volume) use a micro-founded model of the UK economy to give us a broad feel for what drives potential output in the 'flexible-price' economy. Their model departs from standard real business cycle models because firms are allowed to vary the intensity of utilisation of their existing capital stock, at the cost of wearing it down more quickly. The effect is to allow firms more leeway in investment; they can bunch investment decisions because they know that the capital they do have can be made to work harder.

Following a persistent series of total factor productivity (TFP) shocks, the authors show how variable capital utilisation makes investment less correlated with current output and more correlated with future output; investment anticipates underlying productivity improvements. The 'speeding up' of investment also reflects the fact that this is a model of an open economy and capital can be borrowed from abroad.

They further find that, in either of the standard real business cycle or varying capacity utilisation models, the difference between the real interest rate of the flexible-price artificial economies and the observed actual real interest rate seems to play a statistically significant role in explaining UK inflation and output outcomes. These results demonstrate that the flexible-price behaviour of important variables – not just investment but also the real rate of interest – can differ from the smooth processes typically assumed by popular detrending methods, and it can be of interest to monetary policy to explore why.

1.3 What makes a good model for analysing the transmission mechanism?

With the aim of deriving the policy implications of shocks, central banks have long been active in the production of models of the transmission mechanism. These models vary in scope, ranging from single-variable indicators to large-scale macroeconometric models. Whatever the nature of the model of the transmission mechanism, three criteria are traditionally held important in determining a good model for monetary policy purposes.

(a) The model should forecast inflation and output, accurately and robustly, in the face of structural breaks (Clements and Hendry, 1998, 1999).
(b) The model should accurately identify and be correctly conditional on policy reactions, making it immune to the Lucas critique (Bernanke and Mihov, 1998; Banerjee, Hendry and Mizon, 1996).
(c) The model should be estimated on and refer to data that are available and reliable.

Deriving working models that fulfil all three criteria satisfactorily in practice has proved difficult. Later on, we will therefore suggest that transparency has been an additional, important criterion for a good central bank model.

History is cluttered with examples of unreliable forecasts of macroeconomic variables, with such failures afflicting many types of models (Zarnowitz, 1992). The problem is that, although we acknowledge that a model must predict well, how can we assess this predictive capacity when constructing it?

It would seem sensible to choose models that display a good fit and satisfy all the necessary diagnostics indicating a robust specification. But problems arise when the structural breaks – defined as 'permanent large shifts' in the transmission mechanism 'occurring intermittently' (Banerjee, Hendry and Mizon, 1996) – are prevalent. The presence of structural breaks implies that a model passing all these tests on past data need not necessarily forecast the future well (Hendry and Doornik, 1997; Clements and Hendry, 1998, 1999).

Boyd and Smith's paper (chapter 4 in this volume) asks why we have failed to find reliable estimates of the monetary transmission mechanism. The authors' approach is novel. Across a panel data set of annual observations on 60 developing countries, the authors estimate a simple model of the monetary transmission mechanism comprising an IS curve, a trade balance equation, a purchasing parity equation for the exchange rate and a Phillips curve. Their purpose was *not*, however, to produce an average estimate of the transmission mechanism across these countries, but rather to see what we can learn from the wide cross-country differences in the estimates of the parameters.

By comparing the aggregate panel estimates with individual country-specific estimates, the authors show that the dispersion in parameter values is very broad across the countries they sample. They conclude that this diversity seems too large to be attributed to international differences in the transmission mechanism; the process of estimation itself seems to impart biases. What could be the root cause here? The authors suggest that errors in measuring unobservable variables, such as potential output, largely explain why our estimates of the transmission mechanism are so unreliable. These biases appear to vary so much across countries that we should adapt the data, theory and techniques to fit the country-specific circumstances if we are to improve the performance of our estimated models of the transmission mechanism. In their words, more 'tender loving care' is needed.

Structural breaks create additional hazards if they are related to monetary policy. Inaccurate policy advice can be produced from models that are incorrectly conditional on policy reactions; that is, models that do not satisfy condition (b). With this motivation, Muscatelli and Trecroci (chapter 5) look for recent shifts in the transmission mechanism in the USA and the UK that have been caused by changes in monetary policy institutions. More credible monetary policy frameworks in the United Kingdom and the United States may have lowered inflation and, at the same time, made it less volatile. In turn, a more predictable environment may have encouraged firms and workers to establish price and wage contracts in nominal terms. Stickier nominal prices would have slowed down the pass-through of nominal shocks, such as exchange rate changes (Taylor, 2000). Causality could also run in the opposite direction. Changes in the monetary transmission mechanism, such as lower and less volatile global inflation expectations, may have reduced the output cost of anti-inflationary policies in individual countries. These two hypotheses are not necessarily exclusive; an improvement in one country's monetary policy framework may have made it easier for its trading partners to stabilise their own domestic inflation, and vice versa.

Working with postwar US and UK data, Muscatelli and Trecroci test whether changes in monetary policy setting have led to changes in an aggregate demand relationship that determines output and an aggregate supply relationship that determines inflation.[16] They find that, after 1984, it is the inflation equation and not the aggregate demand relationship that appears to have been affected by policy changes in the United States. Both equations seem to have been affected in the United Kingdom over the same period. To estimate the scale and timing of any shift in policy institutions that may have caused this change in the transmission mechanism, the authors estimate a Bayesian (and hence time-varying) vector autoregression (VAR) of the transmission of interest rates, inflation and output gap shocks onto the variables themselves.[17] These time-varying estimates of the responses of the interest rate to shocks suggest that changes in policy reaction function took place in the 1990s for both countries. The responses of UK and US inflation to a shock in the output gap and of UK output to the interest rate shock also seem to have been heightened, albeit gradually, in line with their preliminary findings. These estimates of the interaction between the transmission mechanism and monetary policy in recent years suggest that the Lucas critique is still relevant to estimates of the transmission mechanism, even if it is less visible when inflation is low and stable.

The final requirement of a good model – to be intimate with reliable data – could mean trying to avoid too many unverifiable or unstable parameter calibrations that transport us from the data to the model's forecasts. One example of a family of models that has been well directed to its available data set is the models belonging to the monetary framework of the International Monetary Fund (IMF) (Polak, 1998).[18] The objective of the IMF framework for model design was to forecast what changes in a nation's current economic environment (especially those arising from exports and bank credit) imply for the sustainability of the balance of trade. The approach has proved remarkably influential and durable. From its inception in the 1950s to this day, it continues to perform a key role in the analysis that builds up to the conditionality of IMF borrowing in many countries on IMF programmes.

As Polak (1998) explains, its success may owe much to its simplicity. At the centre of the model is a simple structure that avoids the need to rely excessively on econometric estimates of coefficients from poor-quality data or on theoretical assumptions that may not be valid. To implement the framework, it is not necessary to estimate much more than the nominal income elasticity of the demand for money and parameters for the export and import equations.[19] Inflation and GDP are determined elsewhere but, given the environment to which the framework is applied, the risk of inconsistency is not great. Most of the countries with IMF programmes have fixed exchange rates; prices are therefore determined from abroad. Also, with nominal prices flexible, real output is often dominated by supply-side developments.

1.4 How can these difficulties in modelling the transmission mechanism affect monetary policy in practice?

Kazuo Ueda (chapter 6 in this volume) draws from his experience as a member of the Bank of Japan's policy board to provide a vivid real-world example of the difficulties in understanding the transmission mechanism. From 1998 to 2000, the Japanese economy was in recession, with the overnight call interest rate below 0.5% and prices falling or close to falling. The Bank of Japan lowered the policy rate very close to zero in February 1999; shortly afterwards it announced its commitment to keep the rate at these levels until deflation concerns were alleviated. Ueda's paper analyses this zero interest rate policy, commonly called the ZIRP.

Ueda suggests that the ZIRP had some favourable effect at the time; by committing to a fixed short-term rate for some length of time, the Bank of Japan was able to hold down longer rates and reduce unhelpful uncertainties at the longer end of the yield curve. To have fully exploited the benefits of a ZIRP, however, it would have been necessary to formalise the conditions for terminating it. These conditions could then be made explicit as part of the monetary signal. The conditions for ending the ZIRP could only have been derived if it were possible to assess, with reasonable accuracy, how the transmission mechanism was evolving at that time. The continuing financial crisis and the widespread threat of bankruptcy made it difficult to say confidently where potential output was and where it was heading. Another key parameter, the sensitivity of consumption and investment to the real interest rate, may also have shifted from previous estimates as a result of the credit crunch. Ueda's example illustrates just how serious the measurement difficulties that commonly bedevil monetary policy implementation can, on occasion, become.

Nelson's discussion in chapter 7 of the recent and historical difficulties in understanding the UK monetary transmission mechanism makes an interesting comparison with Ueda's contribution. Other commentators have suggested that the combination of high inflation and high unemployment in the UK from 1965 to 1975 came about because, in policy circles, the view was that permanent improvements in the level of employment could be achieved at the cost of higher rates of inflation. Nelson disagrees. His review of the historical debate shows that influential policy-makers at that time were sceptical of such a systematic trade-off in the UK Phillips curve and became even more so by the early 1970s when evidence against an exploitable trade-off accumulated. He argues that the consensus centred on two views: the coefficient of the output gap in the Phillips curve was too *small* to support such a trade-off, and monetary policy was not very important in determining inflation.[20] In the mid-1970s, these premises led to a situation in which the primary task delegated to monetary policy was to maintain aggregate demand and employment, with the control of inflation left to prices and incomes policies.

Later, the importance of monetary policy in determining UK inflation became ever more firmly established. In 1997, it was enshrined in the instrument-independence of the Bank of England. In parallel, research was redirected towards a better understanding of the transmission of monetary policy to inflation. Perhaps the domestic channels of monetary policy could be best traced through a structural model. In a closed economy setting, that structure would connect the short-run real interest rates to aggregate demand in an IS curve, and then aggregate demand to inflation via the GDP output gap in a Phillips curve.