The Endogeneity of Money:
Empirical Evidence

Peter Howells
University of the West of England

For many years, the endogenous nature of the money supply has been a cornerstone of post-Keynesian economics. In this paper we survey the empirical work which has been done on both the ‘core’ thesis – that loans create deposits – and on peripheral questions such as the origin of the demand for loans, the reconciliation of the demand for money with the loan-created supply and the accommodationist/structuralist debate.

The originality of the paper lies in its demonstration that while post-Keynesians may have thought they were fighting in heroic isolation, most economists involved with the real world of monetary policy-making in practice took much the same view. The consequence is that we can find empirical investigations of issues relating to the endogeneity in a wide range of locations.

*JEL codes:* E50

peter.howells@uwe.ac.uk
1. INTRODUCTION

Any survey of empirical work on the endogeneity of money faces a fundamental problem of where to draw the line. Take the easy case first. We could confine our attention to works which select themselves because their author(s) present them as such. Alternatively, we could use the fact that ‘Endogenous money theory is one of the main cornerstones of Post Keynesian economics’ (Fontana, 2003 p.291)\(^1\) to select on the basis of work which has been published in Post Keynesian, or otherwise sympathetic, contexts. Either approach would draw the line in a broadly similar position.

The problem with this approach is that monetary policy is inevitably pragmatic. Policy must confront what is, even if macroeconomic textbooks continue with the fiction that central banks target the money stock directly (exploiting a mechanical relationship between bank reserves and deposits) and that monetary policy ‘shocks’ must always work through real-balance effects.\(^2\) By contrast, we know that central banks set the rate of interest and allow reserves and deposits to be demand-determined, because they have been telling us so for many years:

…in the real world banks extend credit, creating deposits in the process, and look for the reserves later (Holmes, 1969, p.73).

And more recently:

\(^1\) It might be more accurate to say ‘Endogenous money (as understood in the sense of the next sentence)…’, since Palley (2002) lists ten ways in which the term ‘endogeneity’ may be applied to the money supply and argues that up to eight of these would be consistent with neoclassical treatments of money.

\(^2\) Recent examples include Dornbusch, Fischer and Startz (2003), Mankiw (2003), Burda and Wyplosz (2005). Less excusable is Mishkin (2004) which is offered to readers as a specialist guide to money and banking. Reading these texts, one is reminded of Charles Goodhart’s denunciation of the base-multiplier model of money supply determination as ‘…such an incomplete way of describing the process of the determination of the stock of money that it amounts to misinstruction’. Extraordinary as it may seem, Goodhart wrote those words more than 20 years ago. (Goodhart, 1984, p.188).
In the United Kingdom, money is endogenous – the Bank supplies base money on demand at its prevailing interest rate and broad money is created by the banking system’ (King, 1994 p.264)³

The same message comes from the ‘new consensus view on monetary policy and its recognition of Taylor-rules as a more accurate characterisation of how central banks actually work than any focus on monetary aggregates. The consequence of this is that much empirical work that takes place in central banks and other policy-focused circles is, of necessity, exploring the empirics of endogenous money even if the results are presented as part of mainstream economics with no reference to radical or heterodox origins. Maybe our line should be drawn to encompass any work that recognises that the central bank sets the rate of interest and all else is market-determined. Inevitably, this enlarges the field considerably.

Finally, and closely-related to the pragmatism of policy research, we should recognise that in the UK at least, even when monetary aggregates were of some concern to policy-makers (roughly the period from 1968 to 1985, though with varying degrees of enthusiasm), analysis of the money stock focused upon changes (rather than stocks) and explained these changes through the ‘flow of funds identity’ whereby the flow of new money (on the LHS) was ‘explained’ by its credit counterparts (on the RHS). This is admitting at the very least that the quantity of money is best seen as being determined by the supply of bank credit even if we need to look further for the source of the demand for bank credit in order to clinch the argument that the new money is endogenously-determined. On this view, we should be alert to the possibility that what passes for some very mainstream empirical work in the 1980s, and even earlier, was also in effect shedding some (unintentional) light on the second link in the endogenous money chain: loans create deposits. Taking this approach draws the boundary more widely.

For the purposes of this paper we take the ‘core’ of the endogeneity hypothesis to comprise two causal links: loans depend upon economic activity (broadly defined)⁴ and those loans create deposits. In section 2, we concentrate on tests of the former and in section 3 we look at the loan-deposit link. In both cases we draw the boundaries

³ At the time of their remarks, Alan Holmes was former NYFRB senior vice president while Mervyn King was deputy governor of the Bank of England.
⁴ We return to the meaning of “economic activity” later.
fairly widely. In section 4 we look at the empirical work which has been done on what we shall call ‘secondary’ issues such as the accommodationist/structuralist debate, the reconciliation of the demand for credit with the demand for money, and the link between central bank rate and market interest rates. This arrangement has the incidental effect that we are able to discuss the studies in broadly chronological order. In section 5 we summarise and conclude.

2. THE DEMAND FOR BANK LOANS

Although the endogeneity of the money supply was recognised years ago and had powerful supporters in the not so distant past (e.g. Kaldor 1970, 1982, 1985; Kaldor and Trevithick, 1981; Davidson and Weintraub, 1973) it is Basil Moore who did most to confront the monetarist revival of the 1980s. His book, *Horizontalists and Verticalists* (1988) remains the most substantial theoretical and empirical treatise on the subject, though not now, of course, without its critics.

In that book, Moore set out the core of the hypothesis as we know it, including the assertion that the demand for credit has its origins in the production decisions of firms. In chapter 9 this proposition is subject to empirical testing. The dependent variable is the flow of new bank lending to US industrial and commercial companies (ICCs) 1965(1)-1979(4). The estimation follows the same approach, but covers a slightly different time period, as that used in earlier papers written jointly with Andrew Threadgold. The most accessible was published in *Economica* in 1985 but this had an antecedent in a Bank of England working paper (Moore and Threadgold, 1980). What is remarkable about all three is the theoretical basis which Moore gives for the estimated model. ‘It stems from the recognition that the production process takes time, so that production costs are normally incurred prior to the receipt of sales proceeds.’ (Moore and Threadgold, 1985, p.67). Here we have an explicit test of the hypothesis that the demand for loans (and ultimately the creation of deposits) depends upon the ‘state of trade’ meaning in this case firms’ production plans. If anything happens to increase firms operating costs – an increase in input prices, the decision to increase output, a change in corporate tax rates, *inter alia* – then firms will require more working capital and this will be met by an increase in bank loans.

---

5 e.g. Schumpeter (1911), Wicksell (1898).
Bearing in mind the different data sets, the results are remarkably similar in all three cases. Using the preferred equation from the 1980 paper we have:

\[
LDI_t = 131.6 - 0.727\Delta W_t - 1.162\Delta IB_t - 0.472S_t - 0.932\Delta T_t - 475.4RT - 16.58r
\]

where \( LDI \) is the quarterly flow of new bank lending to ICCs; \( W \) is the wage bill; \( IB \) is the import bill; \( S \) is stock building (additions to inventories); \( T \) is ICCs’ tax bill; \( RT \) is a variable to capture the effect of ‘round tripping’ - periods when ICCs could earn more on deposit than the cost of borrowing, and \( r \) is the real own rate of interest on loans. With the exception of the latter, all variables are in nominal terms. This equation explained about 81 per cent of the flow of new lending to UK ICCs. For the US the figure 75 was per cent (Moore, 1988 p.228).

As a group, these studies appeared to establish what became, for some years at least, two fundamental precepts of the Post Keynesian view of money supply determination. The first is that it was prior changes in firms operating conditions (the ‘state of trade’) that was principally responsible for changes in the flow of new loans (and therefore the rate of deposit creation). This had the effect of establishing firms’ demand for credit at the centre of the money supply process, a practice taken to even greater extremes by the ‘circuitist’ approach to money and credit (see e.g. Graziani 1995). The second was that the interest-elasticity of the demand for credit was low. The latter was important since it purported to show that the central bank’s ability to control credit and money growth was extremely limited. The only instrument at its disposal was a short-term nominal interest rate, and even where changes could be made to have a predictable effect on the real rate, this made little difference to the flow of new loans.

As we hinted in our introduction, however, it was not only economists with a particular point to make about endogenous money who were anxious to explore the behaviour of bank lending in the late 1970s. Recall the context. 1971 had seen the introduction of the ‘Competition and Credit Control’ measures, a package which swept away the long-standing use of administrative controls over money and credit and promised, in future, that the Bank of England would rely solely upon changes in minimum lending rate as its policy instrument. Subsequent events derailed these

---

\(^6\) There was also a study using South African data (see Moore and Smit, 1986) with broadly similar results.
hopes and saw the interest instrument reinforced by the supplementary special deposit scheme (the ‘corset’) intermittently from December 1973. Notice, however, that the SSD scheme was designed in such a way as to penalise banks who allowed ‘excess’ growth in their interest bearing eligible liabilities (‘IBELS’) – roughly speaking their interest-bearing deposits. The purpose of this was to discourage banks bidding for deposits and thus to help drive a wedge between lending and deposit rates. This in turn recognised the point (later publicised by Sprenkle and Miller (1981)) that the demand for both money and credit depended, _inter alia_, on the _spread_ between these two rates. As the spread approaches zero the demand for money and credit tend to infinity.

Later, in 1981, following an extensive review of monetary control arrangements, the primacy of interest rates as the policy instrument was restated, along with the mechanism whereby it was thought to influence money and credit growth. A rise in rates increases the cost of borrowing and thus slows the rate of deposit creation. If, at the same time, lending rates can be raised _relative to_ deposit rates then so much the better since there is then a secondary effect as agents switch from money to non-money assets and lower the cost of non-bank credit relative to bank loans.

In this environment, it is not surprising that policy makers, even from the most orthodox background, would be interested in the demand for bank loans and the late-70s and early-80s saw a flurry of studies, some of which approached the demand for bank credit as a portfolio decision, rather like studies of the demand for money, where the decision to take out a bank loan was considered the equivalent of the purchase of a negative asset, but others, as we see in a moment, certainly recognized the importance of economic activity, in levels and changes.

In the UK, HM Treasury, the National Institute of Economic and Social Research (NIESR) and the Bank of England all had well-developed models of the demand for bank lending and their main characteristics were compared by Cuthbertson and Foster (1982). The Bank of England model was essentially similar

---

7 Examples include Goldfeld (1966) and Jaffée (1971) who focus on utility maximisation within firms’ balance sheet constraints; Méliéz and Pardue (1973) and Méliéz (1975) who take a similar approach to households.

8 Cuthbertson himself went on to develop a more general model of the demand for bank loans by ICCs (Cuthbertson, 1985) in which the HMT and NIESR (but not the Bank of England) models were nested. Later, with the help of J Slow (Cuthbertson and Slow, 1990), he estimated an error correction model for bank advances to ICCs in which changes in the wage and import bill variables were correctly
to the model estimated by Moore and Threadgold (1980) and carried through Moore’s subsequent work: the dependant variable was the flow of nominal lending to ICCs. Firms’ costs were important and the there was a low elasticity on the real cost of borrowing. By contrast, the HMT model was expressed wholly in real terms (except the interest rate) and sought to explain the stock of bank loans while the NIESR model focused, like the Bank of England, on flows but in real terms (but again with nominal interest rates). Only the Bank of England refers explicitly to firms’ costs; the NIESR model had the change in manufacturing output as its activity variable while the HMT model had (the level of) real GDP. Given these differences, it is difficult to compare the models, though if we are prepared to treat real GDP as a proxy for firms’ output then all three find that the level or change in firms’ output is significant. The Bank model, because it is estimated in nominal terms, in effect adds the cost of producing that output as an important variable. The notable differences arise over interest rates, though again direct comparison is difficult. As we have seen the Bank of England/Moore approach gives the own real rate a very low elasticity. The NIESR model differed by treating (nominal) interest rates as a spread (between the return on assets and the cost of borrowing). The results suggested that in the short-run bank lending was highly sensitive but from a policy point of view there was the problem of uncertainty regarding the behaviour of the spread in response to a change in the official rate. If, for example, the rate on assets changed more rapidly than banks repriced their loans in response to an increase in the official rate, there would be the perverse outcome that money and credit expanded. The HMT model, with a (single) nominal rate on bank lending, also showed a negative elasticity, intermediate between the NIESR and Bank of England/Moore and Threadgold equation.

In the 1985 paper Moore and Threadgold claimed that their study was taking ‘…quite a different approach…’ from this earlier work. The basis of the claim was that they recognised explicitly the role of production and the time that it took, while the earlier studies were backed by little theory and involved a rather ad hoc selection of variables (Moore and Threadgold, 1985, p.67). There is some truth in that, but even so it is clear that by the mid-1980s there was a well-established tradition of explaining the demand for bank lending primarily by reference to the demand from

---

9 The genuinely fundamental differences lie between the Moore/NIESR/HMT approach and the studies referred to in footnote 7 above.
ICCs (rather than a broader range of agents) and that this tradition was accepted in orthodox circles. Furthermore, while the HMT and NIESR approaches differed in several details from that of Moore and Threadgold, these can be exaggerated. For example, if we treat GDP and the volume of firms’ output as closely related, then it was widely accepted that both the level and change in real bank lending was significantly influenced by ICCs’ output plans. If we wanted to explain the behaviour of nominal lending then it was legitimate also to consider firms’ costs (at least this seemed to be proven for the flow of nominal lending) and thus to support the view that changes in the money stock were traceable to firms production and costs. Any major differences of opinion concerned the role of interest rates, and this was important because of the emerging consensus that a short-term interest rate, set by the central bank, was the only plausible policy instrument.

We turn now to more recent development of this specific theme – the origin of loan demand – and look at work which is to some degree critical of the two propositions which are commonly associated with the endogeneity hypothesis: that it is firms production plans that are central to the demand for credit and that interest rates have little impact on this demand. We start with the latter.

We have already seen that studies of the demand for credit which otherwise had a number of features in common, differed over the role of interest rates. Furthermore we have also seen that they formulated the interest variable in a variety of different ways: in nominal terms in HMT/NIESR models and in real terms in Moore/Threadgold. Furthermore, the NIESR model had a spread term while the other approaches preferred a ‘simple’ own rate on the cost of borrowing. Before going further, we should note not only the obvious – that the distinction between real and nominal and the possible relevance of relative interest rates, opens up the possibility of virtually unlimited formulations for empirical purposes – but also the fact that the Post Keynesian literature on endogenous money has itself generated a good deal of theoretical controversy on what interest rates are ‘relevant’, much of it associated with the structuralist/accommodationist debate. But the picture is more complex even than this. Not only can one advance different theoretical conceptions of the relevant interest rate and then select from a wide range of different rates for empirical purposes, but when it comes to real rates, judgements have to be made about how to

10 Examples include Palley (1991, 1994), Pollin (1991); Wray (1992); Hewitson (1995); Dow (1996); Howells and Biefang-Frisancho Mariscal (2002); Fontana (2003). And there are others.
model inflation expectations and, in the UK especially, *lending* rates have also to be ‘estimated’ because of UK banks’ reluctance to divulge actual rates on grounds of commercial sensitivity.

The difficulties this can cause was stressed by Hewitson (1997). In the Moore/Threadgold studies the real rate of interest, was the real rate charged on bank loans. This was constructed as the official rate set by the central bank plus a two per cent mark-up, then adjusted for inflation expectations by using a proxy constructed by the Bank of England. Both steps involve untested assumptions. In the case of the mark-up, one might imagine that profit-maximising banks having difficulty in funding the demand for loans would increase their lending rates and thus the mark-up. This of course shows why interest rate behaviour is related to the accommodationist/structuralist debate. If the central bank always ‘accommodates’ fully, why should banks have difficulty in funding new loans? But if one believes that agents, including banks, experience different degrees of liquidity preference in response to different conditions (Bibow, 1998; Dow and Dow, 1995; Dow, 1996; Wray, 1995) then it is quite inappropriate to assume a constant mark-up.

On the empirical side, once again and for similar reasons, the relevant work has been carried out within and beyond a Post Keynesian framework. We come back to the point that if the only plausible instrument of policy is an official rate set by the central bank, then the responsiveness of market rates to any change in the official rate is of widespread and urgent interest. Thus studies at the Bank of England include Dale (1993) and Dale and Haldane (1993). The ‘stickiness’ of market interest rates is also a matter of wider public policy, bearing as it does on the degree of competition in the banking sector and the price that people are paying for financial services and products. This was the motivation behind Shelagh Heffernan’s two studies (1993 and 1997). Within a Post Keynesian framework, and explicitly concerned with the demand for money and the demand for credit there is the evidence (for the UK) in Biefang-Frisancho Mariscal and Howells (2002). What all of these studies show is that there is considerable elasticity in the response of market rates to a change in the official rate. First of all, we need to distinguish a short-run response from the ultimate equilibrium position which may or may not leave relativities unchanged. Even if it does, short-run disequilibrium may be quite sufficiently large and long-lasting to have some effect on the demand for different types of credit. We noted above that it is difficult to get explicit data on bank lending rates, especially to ICCs and therefore
tests of bank loan rate behaviour do not feature largely in the studies above. However, there is much evidence that suggests a constant mark-up is at best a simplification. For example, Heffernan (1997) shows that in response to a change in the official rate, the adjustment in the cost of (bank) mortgage loans was only 37 per cent complete after one month, while the adjustment of personal loan rates took much longer. In the UK, we know that many bank loan rates are priced by a mark-up over LIBOR (London Interbank Offer Rate) rather than the official rate itself, though we know little about the mark-up. However, if LIBOR itself shows a varying relationship with the official rate, then we know that assuming a constant mark-up of loan over official rate is unsatisfactory. This was the thinking behind testing the LIBOR – treasury bill rate (TBR) spread in Biefang-Frisancho Marsical and Howells (2002). Using monthly data from 1986 to 2001 in an unrestricted VAR model they found a cointegrating relationship between LIBOR and TBR. In the long-run, LIBOR is about 0.3 per cent above the official rate. But adjustment of LIBOR to a change in the official rate takes about two months, while adjustment of other (deposit and bond) rates takes much longer. On this evidence, it’s clear that in recent years in the UK bank loan rates will diverge from the official rate, every time there is a change, for a period of at least two months. Furthermore, if we think that spreads are important, the change in spreads will last rather longer.

The modelling of inflationary expectations, in Moore and Threadgold also involves an untested assumption, namely that expectations are actually formed in the way modelled by the Bank of England’s proxy. But more seriously, Hewitson argues, the use of a real rate is anyway inappropriate. It is the cashflow arising from the nominal rate which firms have to meet and if they expect inflation to be high or low, they will expect the nominal value of their revenues to increase rapidly or slowly. That said, one might still expect firms’ demand for credit to be inelastic with respect to the nominal rate since firms face bankruptcy if they cannot meet their (nominal) obligations.

Other criticisms identified by Hewitson included simultaneous equation bias arising from the fact that the central bank’s setting of the rate is partly dependant on the behaviour of variables in the loan demand function. Because of the CB’s reaction function, the official interest rate ceases to be exogenous in the statistical sense. Furthermore, estimating in first differences, as Moore and Threadgold had done, meant that their results contained no useful information about long-run relationships.
Bearing these problems in mind, Hewitson set out to test the demand for credit model using Australian data. The tests were carried out for each of four dependent variables, some monthly and some biannual, ending in 1990, since there was no single Australian series which corresponded exactly to the ICC series used in the UK studies. To avoid the weaknesses in the Moore and Threadgold work, her approach differed in two important respects: firstly the use of an error correction model in order to avoid the spurious regression problem while establishing useful information about long-run relationships; secondly, the use of a nominal interest rate actually charged on bank lending as more appropriate and also as a means of avoiding the simultaneous equation bias that would result by including the central bank’s official rate (+ fixed mark-up).

The estimations were carried out for each of the four variables representing firms’ demand for credit. The results, from a Moore and Threadgold point of view, were disappointing. Sometimes the ECM was insignificant; sometimes the long-run solution derived from the ECM differed from the cointegrating regression from which the ECM was derived; some equations suffered from non-normality of the residuals and, most conspicuously, the wage bill variable was generally insignificant while the rate of interest sometimes was significant, sometimes not. Hewitson’s conclusion listed a number of possible explanations for the results – and suggestions for further testing – but the fact remained that her results were not supportive of the working capital demand for loans and the low interest elasticity.

The fact that changes in ICCs working capital requirements may not provide a very satisfactory explanation for variations in the flow of new bank loans does not, of course, mean that they (and the resulting deposits) are not endogenously determined. Endogeneity requires only that the flow of loans is dependent upon other variables within the system. Instead of concentrating upon firms’ demand for working capital, one might look at a broader based demand for credit. In this case the explanatory variables might include nominal GDP or total final expenditure. A dramatic move towards a broader-based analysis was offered by Howells and Hussein (1999) who experimented with a measure of total transactions (i.e. including transactions in assets and secondhand goods) as the driving force behind the demand for loans. This was motivated by four factors. The first was a recognition that firms’ demand for credit was a diminishing fraction of the total demand which was shifting increasingly towards households. By 1997 60 per cent of total bank and building society lending in
the UK was going to households, compared with only 20 per cent to ICCs (and 20 per cent to financial firms) (Bank of England, 1997). The second was the performance of the monetary sector model developed by Arestis and Biefang-Frisancho Mariscal (1995). This performed quite well when judged by conventional criteria and it was notable for recognising three distinct sources of demand for credit: a household demand for credit (excluding for house purchase), a household demand for mortgage finance and a demand by ICCs. Interestingly, the ICC equation suggested that stockbuilding was a highly significant variable while interest rates were insignificant (broadly supportive of the Moore/Threadgold position). The household demands for credit, however, were influenced by interest rates, specifically by a loan-deposit spread. The model was clearly suggestive of a need to look at broader origins of credit demand.

The third factor was the observation that while total transactions in the UK had been a fairly stable multiple of GDP up to the mid 1970s, the former increased dramatically during the 1980s. As a result, the ratio of transactions to GDP increased from 1.9 in 1976 to 3.1 in 1989. This measure of transactions excluded large value, same day payments between financial firms. Adding those in produces a rise from a ratio of approx. 20 to 55 over the same period. (Howells and Biefang-Frisancho Mariscal, 1992 p. 93). This possibility had been recognised by Keynes in the Treatise (and from Fisher, 1911) both of whom recognised a distinction between ‘income’ (=GDP) transactions and other or ‘financial’ transactions (excluded from GDP). But the distinction was largely forgotten as macroeconomics focused increasingly on income velocity )alone). The significance was pinpointed by Keynes. While income transactions might be closely related to GDP, transactions in secondhand (real or financial) assets:

…need not be, and are not, governed by the volume of current output. The pace at which a circle of financiers, speculators and investors hand round to one another particular pieces of wealth, or title to such, which they are neither producing nor consuming but merely exchanging, bears no definite relation to the rate of current production. The volume of such transactions is subject to very wide and incalculable fluctuations… (Keynes, 1971, V, p.42).

Arestis and Biefang-Frisancho Mariscal (1995, pp.545-6) date the ‘crossover’ in relative shares of bank credit going to UK ICCs and households in 1983.
The final factor motivating an interest in total transactions was the notion that the demand for money (and credit) should depend upon expenditure plans, in general, and not just upon spending upon current output. Where the demand for money was concerned, this was already getting some recognition. For example, Anderson (1993) showed that the boom in mortgage refinancing in the USA had led to an increase in the volume and volatility of financial transactions relative to GDP transactions, and that this had measurable effects upon the demand for M1 deposits. More recently, Palley (1995) and Pollin and Schaberg (1998) had demonstrated that money demand estimates in the USA can be improved by recognising a role for total transactions where the behaviour of the latter is proxied by measures which refer to some part of the property market and to financial activity, two major categories of spending included in total transactions but excluded from conventional measures of GDP. Following this, Howells and Hussein (1997) showed that a total transactions series itself gave better results than either GDP or wealth in an otherwise standard money demand equation.

The first attempt to apply a similar reasoning to the demand for credit was in Howells and Biefang-Frisancho Mariscal (1992). Although the paper was ostensibly about the divergent behaviour of transactions and income velocities, it contained the argument that the demand for credit (and thus the resulting supply of deposits) was heavily influenced by total transactions. It was easily shown that transactions velocity had remained relatively constant. And with total transactions increasing in relation to GDP, this was entirely consistent with the widely-observed fall in income velocity. It was a rather oblique, or initial, formulation of the idea that loans (and deposits) might be endogenously determined but that we should focus upon total, including speculative and property, spending rather than firms’ costs of production.

The Howells and Hussein (1999) paper featured the comparative properties of two estimations, one including GDP and the other including a measure of total transactions as explanatory variables in equations with the flow of real bank lending as the dependant variable. The RHS variables included also a variety of nominal interest rates (on loans, deposits, bonds and on foreign currency) and an inflation term. An ECM (following Banerjee, Hendry and Smith (1986) was estimated using the general to specific procedure. The estimations were based on quarterly data covering the period 1973(1)-1994(4). Changes in both, GDP and total transactions,
were significant at the one per cent level and both models passed the usual diagnostic tests. However, the total transactions model was superior when it came to out of sample forecasts, and when subjected to Akaike and Schwarz Bayesian selection criteria. Interestingly, in the light of Moore and Threadgold’s earlier work, interest rates (domestic, foreign, simple or as spreads) played no significant part. The present state of knowledge does, therefore, seem to suggest that the flow of bank loans is endogenously determined but it is to be explained by a very wide category of spending (rather than by spending of any one group or any particular category of goods and services). It does also seem as if changes in interest rates have rather limited impact.

3. LOANS AND DEPOSITS

The second part of the endogeneity hypothesis, namely that the loans supplied in response to the demand for bank credit create corresponding deposits, has attracted less detailed investigation. No doubt this is because it is less controversial as a result of the banks’ balance sheet identity. The starting point is the recognition that banks’ assets are all, in some form or another, loans. They may be marketable – as commercial or treasury bills – or non-marketable. Even deposits with the central bank are loans (to the public sector). Given this identity, a change in bank loans must be matched by a change in deposits and deposits are the dominant part of the money stock. This is widely recognised in the so called flow of funds identity or ‘credit counterparts’ approach to the analysis of changes in the money stock.

The flow of funds identity can be written with numerous variations but in its simplest form it says:

\[ \Delta D = \Delta BL_p + \Delta BL_g \]  

The change in total bank deposits is the counterpart of changes in total bank lending, i.e. to the non-bank public or to government. For a change in the total stock of money we must add in any change in notes and coin held by the non-bank public and then we can write:

\[ \Delta M = \Delta BL_p + \Delta BL_g + \Delta C_p \]  

Recalling that the government’s total borrowing needs (traditionally called the ‘public sector borrowing requirement’) can be financed in various ways, we can see the
amount of finance provided by banks as a residual, providing whatever is required after all other sources of borrowing have been exhausted. Thus:

\[ PSBR = \Delta G_p + \Delta C_p - \Delta ext + \Delta BL_g \quad \ldots 3 \]

In other words, the budget deficit, adjusted for sales and purchases of foreign currency, must be financed by selling government debt, or by issuing notes and coin or by borrowing from banks.

ADD 4

The FoF approach to the analysis of changes in the money stock has been widespread practice for many years. (Maybe more in the Europe than the US, though it formed the basis of IMF-Bank of England discussion about limiting domestic credit expansion in 1967). Nevertheless, the fact that two sides of an identity must balance is not of itself telling us anything about causality. Indeed, as Cuthbertson (1985b pp.176-7) shows, it is not difficult to rearrange the terms in (3), recall that the monetary base is cash held by the non-bank public plus bank lending to the government in the form of (banks’) reserve assets, and then rewrite (2) so that a change in the base appears on the RHS. In these circumstances, we are back to the base-multiplier model (albeit written in first diofferences) which is the framework normally adopted for the analysis of a money stock exogenously determined by the reserve operations of the central bank.

As we shall see in a moment, making out a case that causality runs from loans to deposits requires some empirical evidence. But it is worth pausing for a moment to consider whether there is not something about the fact that analysis of monetary changes is carried out through the FoF model which disposes researchers to be sympathetic towards the idea of endogeneity. It is a question of the insights that we get from arranging an identity in a particular way. For example, if we write out the base-multiplier model in conventional form, the identity highlights (a) stocks (rather than flows) and (b) suggests that these stocks are fixed unless the central bank decides otherwise. In modern jargon, the ‘default’ setting in the model is a fixed money supply. By contrast, the flow of funds model focuses on flows and encourages us to think of positive flows as the ‘normal’ or ‘default’ case. This is undoubtedly more realistic: central banks, even if they are concerned with monetary aggregates, do not aim at a given stock of money. Their target is the rate of growth and if they do
nothing, in this model, the money supply will continue to expand. It is worth considering that this ‘expansion as the norm’ makes the FoF model an obvious framework for the analysis of an endogenously determined money stock. Furthermore, the model is also very convenient for tracing the effects of the rate of interest when used as the policy instrument. Its relevance to lending to the non-bank public ($\Delta BL_p$) is obvious and the return on government debt is also relevant to debt sales ($\Delta G_p$). While the FoF equation may be just an identity with no behavioural content, it is easy to see how conveniently it can be used to analyse issues surrounding an endogenously determined money supply.

Attempts to establish causality empirically are, inevitably, limited to testing for causality in the statistical sense and this is a good deal more limited than the sense in which we normally use the term. In a statistical context, ‘causal’ means ‘containing information that helps better predict a variable’ (Desai, 1981, p.402). In Moore’s words, ‘What is really being tested is not “$y$ does not cause $x$” but rather ‘an optimal prediction of $x$ does not depend upon $y$’.’ (Moore, 1988 p.151).

In Horizontalists and Verticalists Moore offered a range of Granger and Sims causality tests on four different measures of (US) money, the monetary base and bank loans. The data was monthly and covered, with some variations the period 1974 to 1980. In Moore’s own words:

The evidence presented strongly suggests that unidirectional causality runs from bank lending to each of the four monetary aggregates. Each monetary aggregate has been shown in turn to cause the monetary base unidirectionally. The single exception is the feedback relationship found to exist between the monetary base and M2. (Moore, 1988, pp.162-3. Emphasis added).

A broadly similar investigation by Thomas Palley (1994) albeit for a longer time period, came to similar results.

In 1998 Howells and Hussein followed up both in a study which took quarterly data from the G7 countries for periods which all ended in 1992 or 1993 and began somewhere between 1957 and 1977. The study tested only for the relationship between bank loans and broad money and drew on a number of developments in

---

12 See also Dow (1988) for a methodological discussion of concepts like ‘endogeneity’ and ‘causality’ when applied to the money supply.
econometric methodology after Moore’s 1988 work. For example, the Phillips-Perron (1988) test was used to establish the degree of cointegration between the bank loan and broad money series because of its advantage in allowing for serial correlation and heteroscedasticity. Having established the degree of cointegration, Howells and Hussein then estimated a vector error correction model in which the lagged cointegrating residuals were entered as an explanatory variable and then tested for causality.

The second link in the endogeneity was broadly confirmed across the G7: broad money is ‘Granger-caused’ by loans whatever the reputation and rhetoric of the central bank (Howells and Hussein, 1998 p.337). However, what the results also showed was that there were signs of reverse causality (i.e. from deposits to loans) in each of the G7, most convincingly for Germany and Japan. The presence of bi-directional causality was interpreted by Howells and Hussein as providing some evidence for the debate, which we take up in more detail in the next section, about the relationship between the demand for credit and the demand for money. Put briefly at this point, we can see the problem if we just look back at the FoF identity (4) above. This says that the change in the quantity of money is identical to the sum of credit flows. But money and credit are not the same thing and if we define money, as we do, as a subset of financial assets held by he non-bank public, then (4) raises the question of what has happened to the demand for money? (Not to mention the vast literature devoted to its analysis). As Cuthbertson (1985) says: ‘There is an implicit demand for money in the model [i.e. (4)] but only in equilibrium’ (p.173, emphasis in original). In other words the FoF identity assumes an equilibrium but this raises the question of what sort of mechanism ensures that the flow of new deposits, created by the motives of those who wish to borrow more, will be just equal to the wealth that the mass of the population just wants to hold as extra money balances? Again, briefly, Moore’s position has always been that the willingness to hold the extra deposits (i.e. the demand for money) is irrelevant in a world of endogenous money. Hence, evidence from causality tests that there is feedback from deposits to loans, is evidence that the demand for money does matter and is therefore evidence against Moore’s position.

Just as the Howells/Hussein results were being published, however, other developments on the econometric front were ensuring that the issue would have to be revisited. In 1997, Caporale and Pittis showed that omission of an important variable can result in invalid inference about the causality structure of a bi-variate system
(such as that tested by Moore, 1988, 1989; Palley, 1994 and Howells and Hussein, 1998). However, searching for the role of a third variable by estimating a trivariate vector autoregressive model also has its problems. In Caporale and Howells (2001) the investigators followed the suggestion of Toda and Yamamoto (1995) whose method involved augmenting the correct order of the VAR by the maximal order of integration which characterises the series being used. If $d_{\text{max}}$ represents the latter and $k$ is the correct order, then the $(k+d_{\text{max}})$th-order VAR is estimated and tested in the usual way.

The potential of this approach for exploring the question of causality in an endogenous money context is considerable since it enables us to explore the first link in the endogeneity chain at the same time as we investigate the second. In the present case, for example, we might hypothesise that the omitted variable is the wage bill, or import bill or GDP or, as Caporale and Howells (2001) did, the total transactions variable that performed well in Howells and Hussein (1999). This enabled them to investigate simultaneously the causal impact of total transactions on the flow of new loans and the connection between new loans and new deposits. It also enabled them to explore any direct link between transactions and deposits. The study focused solely on the UK, using quarterly data from 1970 to 1998.

The findings confirmed again the loan $\rightarrow$ deposit link but it also suggested feed back from loans to deposits and, in their interpretation, some role for the demand for money. The novelty in the findings, however, was that while transactions also appeared to cause deposits directly, transactions did not appear to cause loans.

Although the methods are completely different (a demand for loans equation against causality tests) this latter finding must raise some doubt about the inference drawn in Howells and Hussain (1999), namely, that changes in total spending plans drive changes in bank loans.

So far as the second link in the endogeneity chain is concerned, therefore, the present state of empirical knowledge appears to confirm the hypothesis that loans cause deposits. The only controversial aspect of these findings concerns bi-directional causality and the possibility of a feedback from deposits to loans. Howells and Hussein (1998) found widespread evidence for this and it appears again in Caporale and Howells (2001). As we noted above, its relevance to the endogeneity argument is

---

13 For an illustration, see Caporale et al (1998).
not that it threatens the core in anyway, but it relates to what we called in our introduction, one of the secondary questions, namely the question of how the demand for money and the demand for credit interact. It is time now to move on to this and some other secondary issues.

3. REMAINING ISSUES
In this section we look at three more issues which have attracted a degree of empirical investigation. Paramount amongst these, judged by the amount of attention given to it must be the question of why the money supply is endogenous: the accommodationist/structuralist debate.

Before we turn to that, however, we shall tidy up two loose ends by looking at two issues on which we have already touched and which are interrelated, namely the reconciliation of the demand for money with the demand for credit, and the behaviour of market interest rates in response to central bank action.

In the last section we commented on the possibility that causality tests which showed evidence of feedback from deposits to loans might be telling us that a change in the quantity of deposits did not depend solely upon the flow of new loans but might be constrained by the willingness of agents to hold the resulting deposits. This ‘reconciliation problem’ has a long history, though it has not always featured explicitly in accounts of endogenous money. A recent survey of the problem, its history and suggested solutions appears in Howells (2001, §5.8) but we need a brief outline in order to understand better why the causality test evidence and evidence on the behaviour of interest rates may be relevant.

The problem begins with the fact that the demand for bank loans emanates from one set of agents with their own motives (for them it is an income-expenditure decision) while the demand for money emanates from another set with different motives (for them it is a portfolio decision). Granted, the two groups partially overlap but they are not identical. So long as this distinction exists, then there must be a question of how the flow of new deposits, created by a subset of agents with income-expenditure deficits, is to be matched with the population’s desire to arrange their wealth in such a way that they are willing to hold the additional money. And hold it they must (a) in order to satisfy the banks’ balance sheet identity in which loans equal deposits and (b) because money is defined exclusively as deposits held by the non-
bank private sector – in Dennis Robertson’s memorable phrase, ‘All money which is anywhere, must be somewhere’ (Robertson, 1963, p.350).

The possibility that the endogenous production of deposits could differ (*ex ante*) from the demand for them had been raised in 1986 [1992] by Victoria Chick. In what she described as ‘stage 2 banking’ it becomes possible for investment to precede saving because firms can borrow from banks whose loans create the new deposits with which to finance the new investment. Chick’s immediate concern is how exactly to *describe* the extra deposits as a form of saving. But the reference to the demand for them is obvious.

…though the deposits are willingly held there is no actual decision to save. The deposits represent a passive (and grateful) acceptance of means of payment by workers and traders. Some of it will doubtless be used for consumption, some of it saved… [However] while individuals quite happily accepted claims on deposits – acceptability after all is the hallmark of the means of payment – the point on which I wish to insist is that no one actually asks those who subsequently have larger deposits whether the expansion of bank balance sheets was alright with them. (Chick, 1986 [1992] p.200)

But Chick’s target in 1986 was the earlier paper by Kaldor and Trevithick (1981). This had ruled out any problem in connection with the demand for endogenous money on the grounds that there would be an *automatic* application of excess receipts of money to the repayment of overdrafts. Thus, the individual actions of borrowers taking out new loans (or extending existing ones) could threaten an ‘excess’ creation of deposits *ex ante*, but the moment ‘excess’ deposits were recognised they would be devoted by their holders to repaying existing debt. Thus limiting the deposit-creating process – *ex post* – to only those deposits which people were willing to hold.

Note that ‘automatically’ is the keyword. It is a reasonable assumption that those with overdrafts who have receipts in excess of payments will use the excess to reduce their debt and this will (‘automatically’) reduce the quantity of new deposits that are actually created. The problem is - not everyone has an overdraft. And it is not sufficient to argue that some people somewhere (eg virtually all firms) do have overdrafts. Once it is accepted that the first round recipients of ‘new’ money may not wish to hold it, then the problems begin. Some *process* must be triggered by agents as they seek to adjust.
The same issue surfaced in the debate between Charles Goodhart (1989, 1991) and Basil Moore (1991) a few years later and again between Howells (1995, 1997) and Moore (1997). Moore’s position throughout these exchanges is essentially that an equilibrium demand for money makes no sense in a world of endogenous money. In Moore (1991) this was argued on the grounds that if money exists then there must be uncertainty and if there is uncertainty then ‘general equilibrium’ is impossible. This is reinforced by the nature of money which means that it is always accepted as a means of payment. This, argued Moore, shows that there is a permanent and limitless ‘demand’ for money. Agents treat it like a windfall and require no inducements to hold whatever quantity may become available.

However, as Goodhart (1991) points out whatever reservations one may have about general equilibrium this does not entitle one to deny that people have preferences which they will seek to achieve – in this case by swapping money for other assets. And, as Howells (1997) argues, treating ‘acceptability in exchange’ as equivalent to a ‘demand for money’ is to distort the meaning of the latter rather dramatically.

Two areas of empirical work have some bearing on this issue. The first, as we have seen, involves causality tests. If, as seems to be the case, there is some evidence that causality runs from deposits to loans as well as from loans to deposits, then it is hard not to interpret this as some sort of evidence that the demand for money plays a role in the ability of the banking system to create deposits by its response to loan demand. As we saw in the last section, the balance of evidence from causality tests does suggest that the demand for money is relevant.

The second, which we have also touched on, involves the behaviour of interest rates. And this is relevant for two reasons. Firstly, if we accept that there is something in the reconciliation problem, in other words that deposit holders may find their deposits growing more rapidly (or more slowly) than they planned and that they will respond to this by some sort of portfolio reorganisation, then this is bound to affect relative interest rates. Take as an example an ‘excess’ rate of deposit expansion. According to Goodhart (and Howells) agents will periodically exchange their excess deposits for other assets. The price of these assets will rise and their yields will fall. If, for convenience, we talk of ‘bonds’ as the alternative to money then bond rates will fall relative to money’s own rate and also relative to the rate charged on bank loans.

The behaviour of interest rates is relevant also because we recall that the core of the endogeneity thesis involves the proposition that central banks set the rate of interest
and thereafter the quantity of loans (and deposits) is demand determined. Furthermore, central banks are perfectly open in their recognition of an official short-term interest rate as the only feasible instrument of monetary policy. A rise, for example, in the official rate should, it is hoped, reduce the rate of expansion of loans and deposits. But we know from section 2 of this paper that there is considerable debate over the elasticity of the demand for bank loans with respect to the rate of interest and at least some evidence that this may be rather low. In other words, in a conventional diagram, the demand for bank loans is steeply-sloped and a large change in interest rates causes little movement along the quantity axis. But if we think again about relative interest rates, as some of the studies in section 2 do, we can see why the situation may not be so bleak. Look again at the imaginary sequence in the paragraph above; switching between money and bonds causes relative interest rates to change. Now consider the case where the central bank raises the official rate and most interest rates follow roughly in step (quickly at the short end, less directly and more slowly at the long). Suppose now that money’s own rate is sticky. We now have a situation where all rates have risen relative to money’s own rate. This has widened the bond-deposit spread and agents switch from money to bonds, putting a brake on the rise in bond yields. If loan rates adjust by the amount of the official rate change, then (unrestrained) loan rates rise relative to the (moderated) bond rates and another differential widens. Since the issue of bonds is a partial substitute for bank loans and bonds now offer a (relatively) cheaper form of finance, our steep loan demand curve shifts inward and the effectiveness of the official rate change is enhanced. This mechanism and the role of relative interest rates more generally, is discussed in more detail in Arestis and Biefang-Frosancho Mariscal (1995) pp.550-53.

Thus any approach to endogenous money must take interest rates seriously – pertly by virtue of the reconciliation problem, partly because it is central to policy decisions in an endogenous money regime. What happens to relative rates is an empirical question. As early as 1984, Goodhart expressed some scepticism about the authorities’ ability to engineer the sort of relative rate changes we have just outlined as he witnessed the effects of liability management leading to competitive (and ‘unsticky’) deposit rates. This is broadly confirmed in the paper by Howells and Biefang-Frisancho

---

14 ‘It is not that the demand for lending has become less sensitive to changes in relative interest rates. If anything, it has become more so. The problem lies in the increasing inability of the authorities to cause changes in relative rates by changing the level of absolute rates.’ (Goodhart, 1984).
Mariscal (2001) but all of the interest studies referred to in section 2 above\(^\text{15}\) should be read for the light that they shed on the very complex question of relative interest rates and not just for what they tell us about bank mark-ups.

Finally, we need to consider the evidence that has emerged in the course of the accommodationist/structuralist debate, or as it is sometimes caricatured, the question of whether there is a horizontal money supply curve. An altogether more useful way of trying to capture the importance of this debate is to say that it is addressing the question of why the money supply is endogenous. Is it endogenous because, as the accommodationists (e.g. Kaldor, 1970; Moore, 1988; Lavoie, 1992; Rochon, 1999) hold, the central bank willingly makes available any quantity of additional reserves required by banks to validate their lending? Or is it endogenous because, as the structuralists (e.g. Arestis, 1997; Chick, 1983; Dow, 1997; Pollin, 1991; Sawyer, 1996; Wray, 1990) would have it, banks own innovative behaviour enables them either to ‘create’ reserves or to economise on them in such a way that their lending is largely free of any central bank constraint. Fontana (2003) is one of many good summaries of the debate and suggests a diagrammatic representation of the two positions.

We shall see in a moment that some of the ground that we have already covered can be made to bear indirectly on this debate. But the most direct test of these two positions was carried out a few years ago by Pollin (1991). Recall that the accommodationist position assumes that the central bank will always make available the quantity of reserves required by banks in order to support their lending. This may be for political reasons – a desire not to see interest rates rise – or precautionary reasons – a desire to avoid the risk of a banking collapse. If this were the case, Pollin argued, we would expect to see a stable (‘stationary’) relationship between the volume of bank loans and the volume of reserves. On checking this ratio, for the USA, over six NBER cycles from 1953 to 1988, Polin found that the mean increased over time and quite rapidly in later periods. From this, and the behaviour of variances, Pollin drew the conclusion that this favoured the structuralist case – banks appeared to have found ways of increasing their lending per unit of reserves, as if they were reserve constrained.

Another test was suggested by the mechanism encouraging banks to innovate on the reserves front, namely the tendency for market interest rates to rise (relative to the official rate) as the reserve shortage took hold and then to fall as successful innovation

\(^{15}\) See p.8 of this typescript.
eased the shortage. In the presence of structural endogeneity, therefore, interest rate
determination is not a one-way causal process running from the central bank to market
rates: on the contrary, there will be evidence of bi-directional causality. Using Granger-
Sims causality tests and looking at 240 observations on a range of interest rates between
1968 and 1988, Pollin found evidence that two-way causality increased with term. Long
term rates showed signs of a complex interaction with federal funds rate, while short-
term rates appeared to be more directly and unidirectionally ‘caused’ by federal funds
rate. As we said at the beginning of this discussion, the earlier evidence that we reported
on the behaviour of UK interest rate relativities, is not inconsistent with these findings,
though the UK tests were not done for this purpose and were not conducted within a
causality framework.

Although these tests appear to lend considerable support to the structuralist
position, they have not proved definitive. Taking the loan/reserves ratio, for example,
one might argue (see Palley, 2001) that one would expect an increase over time since
banks are profit-seeking institutions and reserves act as a tax on bank outputs. Thus
banks have a continuing incentive to economise on reserves, whatever may be the
position of the central bank and one might imagine that they become more effective in
this as innovations take place.

There is also a problem in the interpretation of the apparent two-way causality
between the Fed and market interest rates. Recall that the bi-directional causality appears
for longer rather than very short rates. How we interpret this should be qualified by how
we think the terms structure of interest rates is determined. Assume that expectations
play some role, i.e. that current long rates embody some expectation of what future short
rates will be. In these circumstances, causality tests of the Granger/Sims types will show
that the future Fed rate is better explained by including a current long rate than without
it.

In more recent years the structuralist/accommodationist debate has become
closely entwined with discussions about the role of liquidity preference, on the part of
both banks and the general public. The connection lies in the fact that if banks are likely
to experience different needs for liquidity depending upon the state of the economy, the
level of perceived risk and so, then all sorts of ratios, not just the loan/deposit ratio, will
be subject to adjustment and banks will ave to find ways of making these adjustments
independent of central bank action. Hence, much of the work that we referred to earlier
(e.g. on the role of the demand for money) has been called upon to support the structuralist cause (see Fontana, 2003, p.297).

4. CONCLUSION

So, after all this effort, what do we know? Firstly, the money supply is endogenous. Leaving aside the writers of macroeconomic textbooks, no one doubts this. Central bank governors tell us it is so and we can see for ourselves that central banks set a rate of interest as the sole instrument of policy, usually amidst a great deal of media attention, and all else is market-determined. There is no point in debating this any more.

But questions of detail remain. Firstly, for example, the question of where precisely within the economic system the demand for loans originates. Traditionally in Post Keynesian economics, and more recently in circuitist theory, the emphasis has been upon firms’ working capital needs. This seems misplaced in the light of the evidence (a) that household demand for bank credit is now (in the USA and UK at least) much larger than firms’ borrowing requirements and (b) that estimates of firms’ loan demand which focuses narrowly upon the wage, import and tax bills do not perform well. Better results are obtained, even for firms’ demand for credit, by including a wider measure of economic activity such as GDP and good estimates of the total demand for credit (including households) can be obtained by recognising that the demand for loans originates with the widest variety of spending plans including those involving secondhand assets. None of this undermines the central hypothesis: the demand for loans originates within the economic system.

Secondly, loans do create deposits. There is ample causality evidence for this. The issue that remains open for further investigation is the significance of the two-way causality. We have seen that this may be some evidence that the demand for money, as well as the demand for loans, plays some part in the money creation process and this would then give some support to the view that we have to think about how the demand for money and the demand for loans are reconciled. One obvious mechanism is largely through portfolio adjustment and (we have seen) this would be supported by what we have also established about changes in interest rates following a change in the central bank’s official rate.

Finally, we can still discuss and investigate the role of the central bank in connection with the supply of reserves. Direct tests of hypotheses associated with the accommodationist and structuralist positions are inconclusive, but could be refined. At
the moment, the structuralist position maybe looks the stronger by virtue of its appeal to evidence about liquidity preference which indirectly lends it some support. But the case is far from closed. It clearly is the case, widely chronicled in the financial press, that central banks go to great lengths to prevent day by day reserve shortages.

References


Jaffee D M (1971) *Credit Rationing and the Commercial Loan Market* (New York: Wiley)


Sawyer M C (1996) ‘Money, finance and interest rates’ in P Arestis (ed) \textit{Keynes, Money and the Open Economy} (Cheltenham: Edward Elgar)


Wicksell K (1936 [1898]) \textit{Interest and Prices: a study of the causes regulating the value of money} (London: Macmillan).


