3 THE EVOLUTION OF BANKING SYSTEMS

Banking evolved from the safe keeping of money.[1] In stylised accounts of the subject people left deposits of a widely recognised monetary commodity (usually a precious metal or 'bullion', such as gold and silver) with a specialist in safe keeping, such as a goldsmith. Initially the deposit was backed 100 per cent by the assumedly safe 'hard' monetary asset. Over time the notes that acknowledged the deposits were used in transactions instead of bullion, while bankers found that they could make loans in their note liabilities instead of tangible gold or silver. By issuing note liabilities without metal backing, the ratio of bullion to total liabilities fell from 100 per cent or more to markedly lower levels. Nowadays the safe monetary asset – the so-called 'monetary base' – is no longer a precious metal, but the legal-tender notes issued by the central bank. But, like gold or silver, legal-tender notes do not pay interest. Because notes are not earning assets, modern banks want to reduce the ratio of cash to their earning assets, in the same way as goldsmiths in embryonic banking.

1 According to the so-called 'de Roover thesis', as a matter of historical fact banking began as a by-product of foreign exchange dealing, with the foreign exchange dealer acting occasionally as a custodian. Interest on deposits was paid from an early stage. See Julius Kirshner (ed.), *Banking, Business and Economic Thought: Selected Studies of Raymond de Roover*, University of Chicago Press, Chicago and London, 1974, pp. 200–201.

Cash in early banking

Early banks often had cash/asset ratios of over 50 per cent. One example is provided by Scottish banking in the middle of the eighteenth century, which is a favourite topic of the advocates of 'free banking'. Indeed, Scottish banking in early modern times was characterised by so-called 'note wars', in which a bank jealous of a rival would encourage business associates to hand over notes and withdraw bullion from that rival so that its bullion would be exhausted![2] But over time banks came to realise that cooperation, as well as competition, had its merits. As well as offering to repay deposits over the counter, banks undertook to make cash payments to third parties on behalf of their customers. So individual A would not need to withdraw £100 of notes from bank X in order to pay them to individual B, who then deposited them at bank Y. Instead bank X would debit £100 from A's account and pay £100 in notes to bank Y, in order that bank Y would credit £100 to B's account. Settling the transaction between A and B via the banks would save legwork and time, particularly if the two banks were located close to each other in a financial metropolis.

So early banking was associated with the establishment of 'note exchanges'. But the physical counting, bundling and transporting of notes remained resource intensive. Real-world payments have always consisted of complicated criss-cross patterns of debits and credits, with most agents having gross incomings and outgoings that are a multiple of the change in their net cash position. Suppose – in our example – that individual C, also with an account at bank Y, wants to make a payment of £100 to individual A. One procedure would be for C to withdraw

2 Charles W. Munn, 'The origins of the Scottish note exchange', *Three Banks Review*, 102, June 1974, pp. 45–60. See, particularly, pp. 50–52.

£100 in notes from bank Y and to pay over the notes to A, who then deposits them with his bank X, at more or less the same time that A is making the £100 payment to B. Alternatively, A could instruct his bank X to pay £100 to B's bank Y, and C could instruct her bank Y to pay £100 to A's bank X. The two banks would see that no movement – indeed, no handling – of notes was necessary at all. At bank X, individual A's account has received a £100 credit and made a £100 payment, and is therefore unchanged, while, at bank Y, B's account has risen by £100 and C's has fallen by the same amount. Transactions to the value of £200 have been carried out, but balance-sheet entries in the banking system have done all the work. Multiplying the £200 by a thousand, a million or a billion times does not affect the principle at work. More explicitly, by adding up all debits and credits for their customers, banks can dispense more or less entirely with the physical handling of notes, and so drastically reduce transactions costs.

But – in our example – what happens if C wanted to make a payment of £120 instead of £100 to A? In that case the business of bank Y's two customers (B and C) would result in instructions to pay £120 in notes and to receive £100, also in notes, so that bank Y must pay £20 in notes (net) over to bank X. The movement of £20 in notes between the two banks would be more economical than the movement of £220 in notes between the two banks and their three customers, but would still be a nuisance. The logical next stage in banking evolution was for a group of banks to form a clearing house, which they both capitalised (in order to pay for the building and infrastructure) and established as an entity where they would maintain deposits. To extend our example, the two banks X and Y know that occasional imbalances in their customers' debits and credits – such as the £20 imbalance referred to above – would occur

from time to time, and so would credit, say, £50 *in notes* to the clearing house. The sum credited by a bank to the clearing house would be the maximum net debit (*in notes*, let it again be emphasised) it expected to arise from its customers' payment instructions, probably plus a small margin for safety. The beauty of the clearing-house arrangement is that, at the end of a particular day's business in which, say, bank Y had a net debit of £35, it does not even have to move notes to other banks, even though strictly its obligation is to pay in notes. Instead bank Y's balance at the clearing house would drop from £50 to £15. If the next day it received net credits of £35, its balance would return to £50. Vast volumes of business can be completed, without any resort to notes as such.

In the historical record the emergence of clearing houses was a gradual process. In England the process was driven by banks' clerks rather than by their proprietors: like so much else in the nitty-gritty of this subject, it was certainly not 'imposed from above'. According to Nevin and Davis's book on *The London Clearing Banks*,

> The first step towards establishing a regular system of clearance was taken by the 'walks' clerks themselves; some time around the mid-eighteenth century, they began to appreciate the advantages to themselves of meeting at a convenient place – usually a coffee house – and exchanging their drafts on each other, settling only the balance in cash. This informal and unauthorised exchange continued for some years until about 1770, when the practice of clearing was accorded official recognition by the private bankers of the City; in 1773 a room was hired for the purpose in the 'Five Bells', Dove Court, off Lombard Street.[3]

3 Edward Nevin and E. W. Davis, *The London Clearing Banks*, Elek, London, 1970, pp. 40–41.

Despite its rudimentary nature, the effect of an organised clearing was to economise on the volume of cash needed in settlement of a given turnover and so to lower the required ratio of cash to assets in banking institutions.

At this stage payment by means of cheques was unusual compared with other types of payment instruction, but over the next hundred years deposit banking via an extensive (and eventually national) branch network became the dominant form of banking business. Once a branch network and a national franchise had been established, payment by cheque took off. A related innovation facilitated these developments. As has been shown, banks could clear by the physical exchange of notes or across accounts in a clearing house. But clearing across accounts at the central bank, the bank which actually issues the notes, is even better. According to a volume issued on the tercentenary of the Bank of England:

> In 1854 joint stock banks in London joined the London Clearing House, and it was agreed that clearing by transfer of Bank of England notes would be abandoned in favour of cheques drawn on bank accounts held at the Bank. Ten years later the Bank of England itself entered the clearing arrangement, and cheques drawn on bankers' accounts at their Bank became considered as paid (i.e., cash).[4]

In late-nineteenth-century Britain most sizeable payments, and the overwhelming majority of transactions by value, were therefore made by cheque. The relevance of commercial banks' clearing arrangements to defining the Bank of England's role is of great importance and will be discussed in more detail shortly.

4 Forrest Capie et al. (eds), *The Future of Central Banking*, Cambridge University Press, Cambridge, 1994, p. 129.

But what happened in the USA, where the Federal Reserve did not exist until 1914?

Were the USA's clearing houses proto-central banks?

The USA has a federal system of government and is of course an enormous country in terms of its land area. Its constitution outlaws the private issue of the legal-tender coin, which is a power reserved to the federal government, but in the early nineteenth century thousands of banks issued notes. The notes circulated on the premise that, when presented to the issuing bank, they could be converted back into coin or 'specie' (i.e. gold or silver). The principle of clearing was well understood, but the multiplicity of note issues and the USA's geographical diversity resulted in a number of regional clearing houses, which contrasted with the undoubted leadership of London in England. A number of banks would participate, with the characteristic pattern being for banks to credit specie or (from 1865) national banknotes at one particular bank (a sort of regional 'central bank'). Inter-bank settlement took place via the resulting accounts at that bank. The celebrated Suffolk Bank system in New England, about which several academic articles have been written, was of this kind.[5]

From time to time confidence in the USA's banks would weaken and banks' note-holders would demand their specie back. Banks could meet these withdrawals either from their own vaults or by taking back some of the bullion left with the clearing-house association. The lower the level of their balance with the clearing

5 Donald J. Mullineaux, 'Competitive monies and the Suffolk Bank system: a contractual perspective', *Southern Economic Journal*, 53(4), 1987, pp. 884–98.

system (i.e., in fact, with the regional 'central bank'), the greater would be the likelihood that individual non-central banks would be overdrawn. (In our example, suppose bank Y's initial deposit with the clearing system was £30. If its customers instructed it to make net cash payments to other banks of £35, bank Y would have been overdrawn by £5.) So a financial crisis and the public's associated large-scale note redemptions would cause increased tension between members of the clearing house.

Although the USA had several large clearing houses, by the middle of the nineteenth century the New York Clearing House Association (NYCHA) was by far the largest and even acted as a kind of reserve clearing house to the regional clearing houses. In a major crisis in 1857 so-called 'country banks' were unable to meet their clearing obligations except by offering their own notes. In order to mitigate the shortage of true legal-tender cash, the NYCHA allowed its members to issue 'clearinghouse loan certificates' against the security of the country banks' notes. These loan certificates were a valid means of settling debts between the NYCHA's member banks. The rationale was that, because the NYCHA's members were confident of their own ultimate solvency, they could economise on specie by granting each other credit. The backing for the loan certificates was a stock of rather unreliable notes issued by the country banks, but – if over time debits and credits between clearing-house members netted out at a very low figure, and the country banks brought their affairs back into good order – that hardly mattered.[6]

Over the next few decades American clearing houses often

6 Richard H. Timberlake, 'The central banking role of clearinghouse associations', *Journal of Money, Credit and Banking*, 16(1), February 1984, pp. 1–15. See particularly pp. 3–4.

issued loan certificates in periods of strain, following the precedent set in 1857. In principle, they were to be called in once specie was again abundant and were viewed as a temporary expedient. But people – including ordinary citizens – regarded them as comparable to specie and they became widely used as a day-to-day currency. In two severe panics, in 1893 and 1907, they were regarded by many contemporaries as a clever expedient, which kept up the effective 'quantity of money' and so offset the deflationary effects of the hoarding of legal-tender coin and national banknotes. In 1907 the total of clearing-house certificates in issue peaked at $88.4 million, compared with the USA's estimated gross national product at the time of over $30 billion and a total of national banknotes outstanding of about $600 million.[7]

As noted by a textbook in the 1930s, the issue of clearing-house loan certificates in the crises of the late nineteenth century '... swept away the necessity of carrying extra till money' and 'by this means the member [of the clearing-house association] was better enabled to meet runs'.[8]

A plausible claim can be made that the USA's clearing houses were proto-central banks and, in that role, helped the banks to reduce the ratio of cash to assets. In the opening years of the twentieth century, however, American bankers and policymakers were unhappy about the performance of their monetary institutions. Although the clearing-house loan certificates facilitated payments and economised on cash, the truth was that the USA

7 The figure for clearing-house loan certificates comes ibid., p. 7. The other numbers are available in a variety of sources.

8 Ray B. Westerfield, *Money, Credit and Banking*, Ronald Press, New York, 1938, pp. 267–8.

lacked an 'elastic' note issue which could be quickly changed in response to banks' needs. American bankers also contrasted the enforced suspension of gold payments in their 1907 crisis with the contemporaneous success of the Bank of England in keeping the pound on the gold standard. Under legislation passed in 1908 a National Monetary Commission was established to investigate the monetary and banking institutions of other countries, and to make recommendations for the USA. The sequel to the commission's work was the establishment of a fully fledged central bank, the Federal Reserve, in 1914.

From Bagehot to Keynes

The flexibility of the US clearing houses in responding to cash runs was impressive and, from today's perspective, constitutes one of the best arguments that a government-sponsored central bank is not an inevitable feature of a modern economy. Nevertheless, the US banking industry did in the end favour their replacement by the kind of central banking arrangements already found in the UK and the rest of Europe. For the bankers it was the so-called 'inelasticity' of the USA's note issue which was its principal weakness. Whereas the Bank of England had a monopoly of the legal-tender note issue, and so could create notes quickly and readily, the US clearing houses could not do that.

The clearing houses did allow their members to operate with lower cash/asset ratios than would otherwise have been the case, but the US system was regarded as inferior to the English. In the late nineteenth century the leading members of the various clearing-house associations were supposed to be particularly solid representatives of US banking and so maintained a high ratio of

cash reserves to deposit liabilities, often above 25 per cent.[9] But in England at the same time banks enjoying the advantages of limited liability, so-called 'joint-stock banks', were operating with cash/deposit ratios of under 10 per cent. In a letter from the governor of the Bank of England to the Chancellor of the Exchequer in the late 1850s, it was pointed out, 'The joint-stock banks of London, judging by their published accounts, have deposits to the amount of £30,000,000. Their capital is not more than £3,000,000, and they have on average £31,000,000 invested in one way or another, leaving only £2,000,000 as a reserve against all this mass of liabilities.'[10]

By implication, their cash/assets ratio was a mere 6 per cent, about a quarter that of similar institutions in the USA. In *Lombard Street*, Bagehot expressed misgivings about the very low ratio of cash to total assets, but fully appreciated the relationship between the banks' cash management practices and their profitability. In his words, 'If they had to keep a much larger part of their reserve in barren cash, their dividends would be reduced, and their present success would become less conspicuous.'[11]

The late nineteenth century saw the continued strengthening of the UK's joint-stock banks, as they perfected the system of inter-bank settlement in a central clearing house based in London

9 'Gillett (1900, pp. 203–4) compares reserve ratios of national banks in the United States to those of British joint-stock banks in the late nineteenth century. He finds reserve ratios of US national banks were more than double their British counterparts.' Charles W. Calomiris, *U S Bank Deregulation in Historical Perspective*, Cambridge University Press, Cambridge and New York, 2000, p. 40. See also Westerfield, op. cit., p. 184.

10 Walter Bagehot, *Lombard Street*, vol. IX in Norman St John-Stevas (ed.). *The Collected Works of Walter Bagehot*, The Economist, London, 1978 (originally published in 1873), p. 176.

11 Ibid., pp. 176–7.

and established national branch networks, often by amalgamation between regional banks. Following the Baring crisis of 1890, the joint-stock banks agreed to publish monthly statements in order that the public would be confident of the quality and safety of their business. An informal understanding was that they should keep their cash/asset ratio above 10 per cent, and in practice that meant a published figure of about 11 per cent. In reality the UK's joint-stock banks, now increasingly known as 'clearing banks', often operated with a somewhat lower ratio and dressed up their monthly numbers so that the published ratio was acceptable to their customers.

When Keynes wrote the newspaper articles that eventually appeared in his 1923 *Tract on Monetary Reform*, the number of London clearing banks had been much reduced by mergers and amalgamations. The five largest London clearing banks dominated English banking, and together with the two big Scottish banks (the Bank of Scotland and the Royal Bank of Scotland) they also dominated British banking. In Keynes's words, the Big Five banks' 'aggregate deposits have always been about nine times their "cash"'. Since the implied 11 per cent ratio is 'generally considered a "safe" proportion, it is bad for a bank's reputation to fall below it, while on the other hand it is bad for its earning power to rise above it'.[12] These arrangements continued with little change throughout the inter-war period, with Keynes giving a further, more detailed description in his 1930 *Treatise on Money*. While his treatments noted that to a large extent banks

12 John Maynard Keynes, *A Tract on Monetary Reform*, vol. IV in Donald Moggridge and Elizabeth Johnson (eds), *The Collected Writings of John Maynard Keynes*, Macmillan for the Royal Economic Society, London and Basingstoke, 1971 (originally published in 1923), p. 142.

operated with stable ratios because of their convenience as rules of thumb, both the *Tract on Monetary Reform* and the *Treatise on Money* alluded to the tension between the banking system's safety (increasing with the cash/assets ratio) and its profitability (decreasing with the cash/assets ratio). Keynes, like Bagehot, saw that any discussion of banking industry structure had to recognise that commercial banks were privately owned organisations with profit as one of their main objectives.

The ability to operate with an apparent cash ratio of 11 per cent, and a true cash ratio of 10 per cent or less, had been facilitated by two insights. The first was that the convertibility of deposits into cash could be protected by holding interest-bearing assets that could be readily sold for cash as well as by the holding of cash itself. The payment of interest on such 'liquid' assets helped profits, while their ready saleability for cash protected depositors. Second, if a distinct institution with the prerogative to issue notes (i.e. a central bank, which was the Bank of England in the UK's case) assumed a responsibility to lend to commercial banks if they ran out of cash, those banks could operate with lower cash/asset ratios than before.[13] Indeed, the key to maintaining deposit convertibility was not merely to have a large holding of idle vault cash, but also to nurture a good relationship with the Bank of England and keep holdings of an assortment of 'liquid assets'. It was understood that such assets could either be sold to the Bank or would serve as collateral for a loan. According to Nevis and Davis in their historical account in *The London Clearing Banks*, 'Improving communications and ready access to head

13 The assistance to the banking system might come in the form of purchases of securities, perhaps from strongly capitalised and liquid banks, and not just in the form of loans to banks. See Bagehot, op. cit., pp. 134–5.

offices, together with re-discounting facilities in the London bill market and the emergence of the Bank of England as "lender of last resort", had resulted in a tendency to work to a minimum of till money.'[14] Indeed, the British system of a small number of large clearing banks, with national branch networks and close connections with a central bank (i.e. the Bank of England) that would occasionally lend to them, 'was to serve as a model for monetary authorities throughout the world'.[15]

Trends in British banking in the second half of the twentieth century

Commercial banks' holdings of liquid assets other than cash improved the trade-off between profitability and depositor safety. For most of the twentieth century the Bank of England therefore paid close attention both to the clearing banks' cash ratio *and* to their 'liquidity ratio' (i.e. ratio of explicitly defined liquid assets to deposits held by non-banks). In the first few years after World War II the cash ratio dropped to 8 per cent, while the liquidity ratio was 40 per cent and banks' assets were dominated by claims on government. In such circumstances it was virtually inconceivable that a run would exhaust banks' cash holdings. A run might do serious damage to banks' initial cash holdings, but they could quickly sell some of their government securities to the Bank of England for cash, and so replenish the cash in their tills and vaults.

Over time the two ratios fell dramatically. By the late 1950s the Bank of England had allowed the liquidity ratio to go down to about 30 per cent, although the institutions specifically charged

14 Nevin and Davis, op. cit., p. 78.
15 Ibid., p. 82

with respecting this ratio – the clearing banks – resented the competition they faced from other credit-granting organisations not subject to ratio control. In the Competition and Credit Control reforms of 1971 the alleged discrimination against the clearers was largely remedied by the setting of a 'reserve assets ratio', applicable to all banks, at 12.5 per cent of sterling deposits. The clearers had to keep a non-interest-bearing balance at the Bank of England, equal to 1.5 per cent of deposits, on top of their required reserve assets, but this had an obvious functional rationale in their clearing activities and was not objectionable to them.

By now competition and risk-taking were intensifying, but British banking seemed still to be working smoothly. Although large-scale retail runs were mentioned in the history books, they no longer figured in the memory of anyone actually working in a British bank. In 1981 both the clearers' 1.5 per cent ratio and the 12.5 per cent reserve assets ratio were scrapped. Instead all banks – whether involved in clearing or not – were to lodge a deposit in 'special non-operational, non-interest-bearing accounts' at the Bank of England equal to 0.5 per cent of so-called 'eligible liabilities' (i.e. non-equity liabilities to agents other than banks and the government). Partly because of the fading collective memory of bank runs, these accounts were seen as serving no purpose in either monetary control or financial supervision and regulation. Instead they were understood to be a special mechanism, in effect a form of tax hypothecation, which gave the Bank of England funds to reinvest in interest-bearing securities and so to generate an income sufficient to cover its costs. The clearers kept a separate balance, over and above the 0.5 per cent, to settle debit and credit balances at the end of each daily clearing, but it was now a very low proportion of their balance sheet totals.

The Bank of England was still concerned about the degree of maturity transformation that the banks were undertaking. (Maturity transformation is the extending of long-term loans against short-term liabilities, including deposits repayable on demand.) The liquidity ratio was history and the reserve asset ratio had been abolished, but in July 1982 the Bank published a paper on 'The measurement of liquidity', showing how individual banks were to calculate (among other things) a 'net cumulative mismatched position'. Bank officials continued to supervise all banks' liquidity until 1998, when the job was transferred to the newly created Financial Services Authority as part of an institutional upheaval at the start of the Blair government. This institutional upheaval led to the transfer of many officials from the Bank of England, with its decades of experience and a fund of central banking know-how, to the FSA, which had yet to find its feet. Some officials at the FSA undoubtedly did appreciate that the structure of assets, and in particular the ratios of cash and liquidity to total assets, was relevant to the integrity of the banking businesses under its supervision. But a fair comment is that official interest in UK banks' ability to withstand a run was sharply less than had been the norm during the twentieth century.[16]

The insouciance towards banks' vulnerability in a run was

16 Several articles appeared in the Bank of England's *Financial Stability Review* between 2000 and 2005 on the UK banking system's liquidity. An article in the December 2000 issue (pp. 93–111), on 'Banking system liquidity: developments and issues', by Graeme Chaplin, Alison Emblow and Ian Michael, opined that 'the extent of maturity transformation at a three-month horizon in the UK banking system seems to be fairly stable through time'. A speech in November 2005 on 'Financial stability: managing liquidity risk in a global system' (pp. 78–84 in the December 2005 *FSR*) by Sir Andrew Large, then Deputy Governor for Financial Stability, correctly identified some of the problems that came out into the open less than two years later, but he left the Bank of England shortly afterwards.

reflected in several developments in the decade leading up to the Northern Rock crisis. Earlier discussion in this chapter established that banks' cash reserves with the Bank of England had a definite functional rationale for the depositing banks themselves. Their cash reserves were both the accounts in which the clearing banks themselves settled their end-of-day imbalances and a backstop for their vault cash, if their vault cash came under attack from a loss of confidence and a retail run. Further, by opening an account at the Bank of England a bank started a relationship with the UK's central bank, which included the possibility of borrowing from it in the appropriate circumstances. Some types of so-called 'bank' did not have an account at the Bank of England and could not appeal to it if they ran out of cash. Indeed, historically, building societies had not maintained accounts at the Bank of England. Instead they 'banked' via the clearing banks, while they had been regulated not by the Bank of England, but by the Registrar of Friendly Societies.[17]

But officialdom seems increasingly to have forgotten that banks' cash reserves at the Bank of England had an operational purpose. Under the terms of the 1998 Bank of England Act and the 2000 Financial Services and Markets Act, all UK 'banks' were required to maintain a non-interest-bearing balance at the Bank of England whether they undertook clearing and payments settlement business or not. Admittedly, the requirement was only 0.15 per cent of eligible liabilities and so was hardly a big threat to their profits. The Treasury subsequently published two consultative papers on what it had come to term 'the cash ratio deposit

17 Jack Revell, *The British Financial System*, Macmillan, London and Basingstoke, 1973, p. 367.

scheme'.[18] The scheme was discussed solely and entirely as a mechanism for covering the Bank of England's costs, and as having no wider value for the British banking system. The two documents seemed to be oblivious to the traditional rationale of a cash reserve at the central bank from the commercial banks' own point of view.

Before its demutualisation in October 1997, Northern Rock had been a mutually owned building society and its direct contacts with the Bank of England were perfunctory. Since 1998, like other British banks, it has kept a non-interest-bearing deposit at the Bank of England. In May 2006 the Bank of England changed the structure of its relationship with the UK's commercial banks in wide-ranging reforms, notably by starting to pay interest on cash reserves separate from the 0.15 per cent cash ratio deposit scheme. The new terms of the Bank of England's relationship with its customer banks were contained in a Red Book, which – in its own words – was 'designed to provide flexible access to central bank money, including in unlimited size against eligible collateral at a penalty rate through' the so-called 'standing lending facility'.[19]

18 The Treasury published two documents – both called *Review of the Cash Ratio Deposit Scheme: Consultation on proposed changes* – in August 2003 and August 2007. In qualification to the statement in the text, the Bank of England was fully aware of the significance of the cash ratio deposit scheme for banks' liquidity management. See *The Framework for the Bank of England's Operations in the Sterling Money Markets* ('the Red Book'), Bank of England, London, March 2008, p. 6.

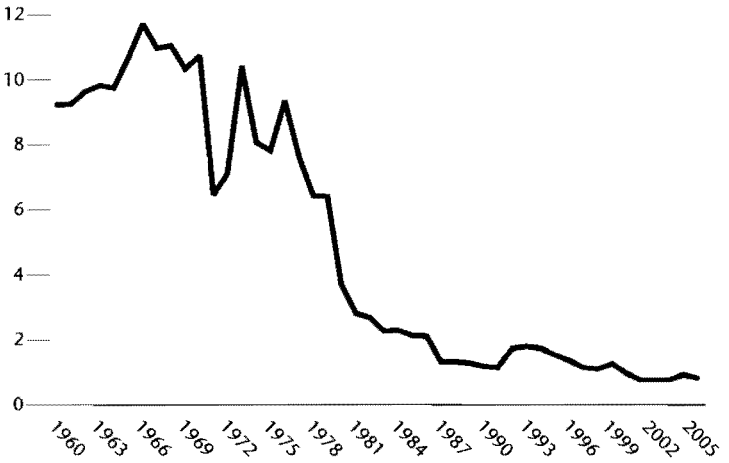
19 'Red Book', p. 7. An important technical detail needs to be mentioned. Advances in computer technology enabled real-time gross settlement (RTGS) to be introduced for large sums in 1996. The change from the settlement of a balance at the end of the day (i.e. as in the daily cheque clearing) to RTGS enabled banks further to economise on cash. Some Bank of England officials have subsequently preferred the phrase 'settlement banks' to 'clearing banks', but 'clearing banks' remains the most common usage and has been retained here. The author understands that the RTGS equipment often broke down in the early years and more primitive systems had to be used as a back-up.

In the summer of 2007 Northern Rock was a participant in the Bank of England's reserve schemes and a member of the list of banks to which a standing facility might be granted. On the face of it, Northern Rock would have been behaving reasonably in expecting the Bank of England to be helpful, or at least 'flexible', if it had trouble financing its assets.

British banks' negligible cash holdings in the early 21st century

The events of August and September 2007 were to show that, in practice, no one in Northern Rock's management or the Bank of England knew precisely what was supposed to happen if Northern Rock lost the confidence of its retail depositors. Nevertheless, for most of Northern Rock's existence as a PLC the resilience of its defences against a retail run was not a big topic in its corporate strategy. Its regional roots and smallness handicapped it in the market for UK retail deposits. Here the clearers – with their national branch networks and the scale that allowed them to enjoy huge 'network economy' advantages in settlement business – were entrenched. But in truth, by the early 21st century the whole of the British banking system had economised on cash to a remarkable extent and, in this respect, taken a cavalier attitude towards funding risk. Cash as a fraction of total sterling liabilities, and even of sight sterling liabilities, had become nugatory by 2005. In January 2006 UK banks' cash ratio deposits were £1,953 million and other balances at the Bank of England (i.e. the balances actively used in settlement of payments business) were £839 million, and their vault cash was £5,417 million. Their total cash resources were therefore £8,209 million. At the same time

Figure 1 **Collapse in UK banks' cash/deposit ratio, 1960–2005**
Ratio of cash reserves to 'sight, time, savings and foreign currency deposits' of UK banks, IMF data



According to IMF data (which have the advantage of continuity over 45 years), the UK banks' cash/deposits ratio fell from over 9% in 1960 to under 1% in 2005.
Source: IMF

their sight liabilities to UK non-banks were £629,892 million and their total sterling liabilities £2,534,494 million. So the ratio of cash to sight liabilities held 'by the British public' was 1.3 per cent and the ratio of cash to all sterling liabilities was 0.3 per cent.[20] In other words, the cash ratio of British banks had dropped to about a thirtieth of what it had been 80 years earlier! Perhaps it is unnecessary to add that the situation in the summer of 2007 – which had changed again because of the introduction of interest-earning

20 Figures are taken from relevant issues of *Financial Statistics*, The Stationery Office, London.

reserves in May 2006 – remained a far cry from the 100 per cent cash reserve ratio found when the idea of banking had been conceived in the late Middle Ages.

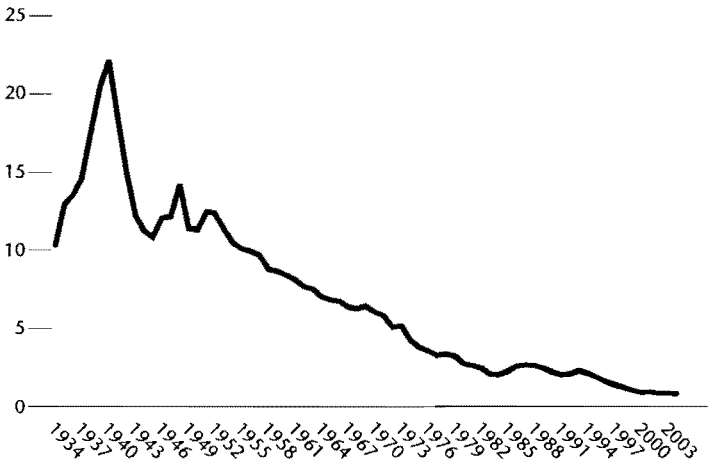
At any rate, the historical review over the last few pages has shown that British banking – which started, like all banking, with a cash/assets ratio of 100 per cent – was able to operate successfully for several years with a cash/assets ratio that was a fraction of 1 per cent. The spectacular reduction in banks' cash holdings had been made possible, among other things, by the help given by the Bank of England to Britain's banks in their balance sheet management. Critically, the Bank of England had been able and willing both to purchase a range of assets from them, and occasionally to lend to them, in order to relieve any cash shortages.

And what about banks in the USA and Europe?

Several books could be written about the cash-holding behaviour of the world's banking systems over the long run. There is room here only for a brief review of developments in the USA and Europe, in order to give a broader and more international perspective.

As explained above, from its inception one aim of the Federal Reserve System was to provide an elastic supply of cash to member banks and so lower US banks' cash/assets ratios. This was indeed an initial result of its creation. A contemporary US textbook on money and banking opined that 'as a constituent of our circulating media' the cash element (coins, government paper money and banknotes) was 'a small and declining proportion', and quoted a calculation by Angell and Ficek that cash had fallen from 18.1 per cent of the total circulating media in 1909 to 7.7 per cent

Figure 2 **US banks' cash/deposit ratios, 1934–2004**
% of deposit liabilities



Source: Federal Deposit Insurance Corporation website

in 1930.[21] But the Great Depression of the early 1930s resulted in thousands of bank failures, and so caused both banks and people to hold more cash relative to their other assets. As a precaution against the return of troubled times, US banks' cash/deposits ratios were back above 20 per cent by the early 1940s. After World War II, however, banks worked together with the Federal Reserve to bring cash/deposits ratios down, with the results shown in Figure 2. The reduction in banks' cash/deposits ratios was helped by depositors' growing preference for time deposits, which by the 1980s paid an attractive interest rate. (Banks did not need to keep

21 Westerfield, op. cit., p. 184.

a cash reserve against time deposits, in contrast to sight deposits for which US officialdom had always demanded a cash reserve, both before and after the establishment of the Federal Reserve.) At any rate, by 2005 the ratio of cash to all deposits in US banking was under 1 per cent, not dissimilar to the figure in the UK.[22]

Until the introduction of the single European currency in 1999, the notion of a 'European banking system' was misplaced. The structures of banking systems varied from nation to nation and did not follow an exactly uniform pattern of development. Cash ratios took widely divergent values, with a compulsory cash ratio set well above banks' true functional needs in some countries.[23] When the euro was established on a scriptural basis in 1999, the European Central Bank began to pay interest on banks' cash reserves with it. This was a revolutionary innovation which – in one quantum leap – meant that Europe's banks were more favourably placed in their cash management arrangements than their counterparts in the USA or the UK. (Whether the payment of interest on reserves is good for the banking system's efficiency is a larger topic[24].) Apart from vault cash, banks in the newly formed single currency area had virtually no non-earning assets whatsoever. The same generalisation applies in the member states of the Eurozone as in the UK and the USA – that advances in central

22 Data are available from several websites, including those of the Federal Reserve and the Federal Deposit Insurance Corporation.

23 For example, Spain's banks in the 1980s were subject to a system of '*coeficientes*', required ratios of government debt to total assets. A senior banking executive is reported to have asked, 'Am I a banker at all? I am not allowed to be one.' T. Burns, 'Hamstrung by siphoning of deposits', *Financial Times*, 3 April 1985.

24 If interest is paid on banks' cash reserves at the central bank, they have less incentive to lend in the inter-bank market. Whether this affects the efficiency of the inter-bank market is moot, but see Tim Congdon and Brandon Davies, 'A simple plan to unclog the interbank market', *Financial Times*, 23 October 2008.

banking have helped banks to lower drastically the ratios of non-earning cash assets to total assets.

Long-run trends in bank solvency

For many decades, writers on monetary theory – and particularly writers on the theory of monetary policymaking – paid considerable attention to the ratios of cash and liquidity to banks' overall balance sheet size. By contrast, the ratio of capital to assets was neglected until the 1980s. One reason is that central banks did not always publicise their views on the desirable level of banks' capital/assets ratio. In the first edition of *The British Financial System*, published in 1973, Revell noted that building societies were subject to regulations on their capital reserves set by the Registrar of Friendly Societies, but for the banks matters were somewhat different. To quote, 'The Bank of England keeps a close watch on the reserve ratios of the bodies under its direct surveillance in the banking system – deposit banks, accepting houses, other secondary banks and discount houses. In all cases it works to certain minimum ratios, although nobody outside the Bank knows what these ratios are.'[25]

Of course, banks' management were cognisant of their capital ratios from internal records and they had to keep shareholders informed in their audited accounts. So – despite the apparent regulatory neglect of the capital side of banks' balance sheets until the last 25 years – researchers have been able to compile data on the long-run behaviour of capital ratios. As with the cash and liquidity ratios, the trend is clear. Whereas in the embryonic phase of banking capital/asset ratios put heavy emphasis on safety

25 Revell, op. cit., p. 105.

and were often over 30 per cent, in the twentieth century and the opening years of the 21st century the ratios fell substantially and with only occasional interruptions to the long-run pattern.[26]

This is not the place for a systematic treatment, but a summary verdict can be offered. (See Table 2 for relevant data.)[27] In the late nineteenth century a capital/assets ratio of over 15 per cent was normal even in the UK, the most advanced financial power of the time. By contrast, in the first half of the twentieth century the leading British banks regarded an appropriate capital/asset ratio as between 7 and 10 per cent. In the second half of the twentieth century the figure had fallen to 5 to 6 per cent. By the end of the century banks in the USA and Europe – which had historically operated on higher capital/asset ratios than their British counterparts – increasingly had the same attitude towards capital adequacy, but bank managements and regulators in these areas were dismayed by the very low capital/asset ratios in Japanese banking. Indeed, the view that the thinness of Japanese banks' capital cover allowed them to undercut their rivals in the offshore banking markets provoked the Anglo-American 'convergence accord' on bank capital in January 1987. The accord developed into the Basel capital rules, which were enforced in all the participating countries, including Japan, to establish a 'level playing field'. As is well known, the central principle of the first set of Basel guidelines was that capital should be not less than

26 The subject is of course vast, but – for example – see p. 124 of Howard Bodenhorn, *A History of Banking in Antebellum America*, Cambridge University Press, Cambridge, 2000. At end-June 1840 the Bank of Charleston had an equity-to-assets ratio of 60.6 per cent and a contingency fund of over 5 per cent of assets as well! It nevertheless earned a return on equity of about 10 per cent.

27 The data used in the table come from p. 149 of M. K. Lewis and K. T. Davis, *Domestic and International Banking*, Phillip Allan, Oxford, 1987.

8 per cent of assets, with equity capital equal to at least half of total capital. The similarity of this principle to the capital/assets ratio of about 5 per cent preferred by Britain's banks is striking. Given the pattern of the preceding international negotiations in which UK officials had been so active, the setting of a 4 per cent minimum may not have been entirely accidental.

Table 2 **Equity capital to total assets of UK and US banks, 1880–1985**

	*UK banks**	*US banks†*
1880	16.8	n/a
1900	12.0	n/a
1914	8.7	18.3
1930	7.2	14.2
1940	5.2	9.1
1950	2.7	6.7
1966	5.3	7.8
1980	5.9	6.8
1985	4.6	6.9

*UK deposit banks 1880–1966, UK clearing bank groups 1980 and 1985
†All member banks of the Federal Reserve system
The low value of the UK ratio in 1950 reflected the high ratio of low-risk government paper in banks' assets after World War II.
Source: See note 27

Like all other British banks, Northern Rock was subject to the Basel rules at its demutualisation. Indeed, references to compliance with the latest developments in the Basel regulatory framework were included in its last published accounts as a quoted PLC, only a few weeks before its collapse.[28] Perhaps it is premature to

28 Also neglected – as mentioned in the text – are the complications arising from banks' issue of bonds and reference capital. Liabilities are deemed, for simplicity, to consist solely of equity capital and deposits.

pass judgement on international banks' manipulation of asset and liability structures over the last decade or so, as they attempted to bypass the Basel constraints by the creation of artificial 'special purpose vehicles' or 'conduits'. Nevertheless, even a cursory examination of banks' annual reports shows that in the last few years actual ratios of equity capital to assets have often been under 3 per cent for a very large number of institutions. They nevertheless met the Basel rules because those rules allow a zero weight (in terms of capital usage) for inter-bank exposures and claims on government, as well as other technical exemptions.

To conclude this section, in the early phase of modern industrialism banks typically had capital/asset ratios of over 30 per cent, but in the middle years of the present decade the 'average' ratio of equity capital to assets (if the phrase has any definite meaning) may have been about 5 per cent and the effective ratio for a surprisingly high number of prominent institutions was little more than 3 per cent.

What do the trends in liquidity and solvency imply for loan margins?

It is now time to bring together the strands of the argument by setting out a table which shows how, for a particular target return on equity, the average return on bank assets varies with different ratios of cash and capital to assets. Table 3 uses the formula developed at the end of Chapter 2 for the determination of banks' average return on assets. A reminder may usefully be inserted that the implicit assumptions in preparing the matrix are the same as they have been throughout this paper, that banks have no loan losses, and that their fee revenues cover the costs of organising the

loans, and of running any deposit collection and money transmission infrastructure.

A target rate of return on capital of 14 per cent has been chosen, as this sort of number would be regarded as appropriate by contemporary UK banks in their internal strategy documents and serves as a reasonable benchmark for discussion.[29] In the very early days of banking – when banking was indeed little different from risky and avaricious moneylending, and the cash ratio was perhaps 80 per cent and the capital/assets ratio 45 per cent – the loan margin had to exceed 30 per cent. In the opening decades of the Industrial Revolution, in such countries as England, Scotland and the USA, a cash ratio of 40 per cent and a capital/asset ratio of 20 per cent would have been commonplace in the banking industry. A loan margin of almost 5 per cent (i.e. 500 basis points) would achieve a return on capital of 14 per cent. In the early decades of the post-war world, with a cash ratio of 5 per cent and a capital/assets ratio of 8 per cent, a loan margin of about 200 basis points would have been consistent with that return on capital. But in the low-ratio banking of the last fifteen years or so, loan margins of 100 basis points or less were compatible – assuming all went well with asset selection and cost control – with high bank profitability.

29 Northern Rock was one of several British banks to have exceeded the 14 per cent figure by a wide margin for many years, until its funding – and so the business itself – collapsed in late 2007. The chief economic commentator of the *Financial Times*, Martin Wolf, protested about the high profitability of banking in a column on 28 November 2007, attributing it to 'sundry explicit and implicit guarantees' from the state. Later in his column he endorsed 'higher capital requirements' for banks.

Table 3 **How banks' loan margins vary with their cash and solvency ratios**

The table shows, with a given target rate of return on capital, how a reduction in banks' average return on assets (i.e. their 'loan margin', more or less) becomes possible as their cash/asset and capita/asset ratios decline. All figures are expressed as a percentage.

P/K *Rate of return on capital*	*C* *Cash ratio*	*K/A* *Capital/assets ratio*	r_b *'Loan margin'*
14	80	45	31.5
14	40	20	4.7
14	12	15	2.4
14	5	8	2.2
14	1	5	0.8
14	1	3	0.5