Andrew Haldane: The contribution of the financial sector – miracle or mirage?

Speech by Mr Andrew Haldane, Executive Director, Financial Stability, of the Bank of England, at the Future of Finance conference, London, 14 July 2010.

This text is taken from a co-authored chapter by Andrew Haldane, Simon Brennan and Vasileios Madouros in "The Future of Finance: The LSE Report" published today by The London School of Economics.

* * *

We would like to thank Stephen Burgess, Melissa Davey, Rob Elder, Perry Francis, Jen Han, Sam Knott, Nick Oulton, Peter Richardson, Jeremy Rowe, Chris Shadforth, Sally Srinivasan and Iain de Weymarn for comments and discussion on earlier drafts, and Alexander Haywood and Laura Wightman for research assistance. The views expressed are those of the authors and not necessarily those of the Bank of England.

1. Introduction

The financial crisis of the past three years has, on any measure, been extremely costly. As in past financial crises, public sector debt seems set to double relative to national income in a number of countries (Reinhart and Rogoff (2009)). And measures of foregone output, now and in the future, put the net present value cost of the crisis at anywhere between one and five times annual world GDP (Haldane (2010)). Either way, the scars from the current crisis seem likely to be felt for a generation.

It is against this backdrop that an intense debate is underway internationally about reform of finance (Goodhart (2010)). Many of the key planks of that debate are covered in other chapters in this volume. Some of these reform measures are extensions or elaborations of existing regulatory initiatives – for example, higher buffers of higher quality capital and liquidity. Others propose a reorientation of existing regulatory apparatus – for example, through counter-cyclical adjustments in prudential policy (Bank of England (2009b), Large (2010)). Others still suggest a root-and-branch restructuring of finance – for example, by limiting the size and/or scope of banking (Kay (2009), Kotlikoff (2010)).

In evaluating these reform proposals, it is clearly important that the on-going benefits of finance are properly weighed alongside the costs of crisis. Doing so requires an understanding and measurement of the contribution made by the financial sector to economic well-being. This is important both for making sense of the past (during which time the role of finance has grown) and for shaping the future (during which it is possible the role of finance may shrink).

While simple in principle, this measurement exercise is far from straightforward in practice. Recent experience makes clear the extent of the problem. In September 2008, the collapse of Lehman Brothers precipitated a chain reaction in financial markets. This brought the financial system, and many of the world's largest institutions, close to the point of collapse. During the fourth quarter of 2008, equity prices of the major global banks fell by around 50% on average, a loss of market value of around $640 billion. As a consequence, world GDP and world trade are estimated to have fallen at an annualised rate of about 6% and 25% respectively in 2008Q4. Banking contributed to a Great Recession on a scale last seen at the time of the Great Depression.

Yet the official statistics on the contribution of the financial sector paint a rather different picture. According to the National Accounts, the nominal gross value-added (GVA) of the financial sector in the UK grew at the fastest pace on record in 2008Q4. As a share of whole-economy output, the direct contribution of the UK financial sector rose to 9% in the last quarter of 2008. Financial corporations' gross operating surplus (GVA less compensation for employees and other taxes on production) increased by £5.0bn to £20bn, also the largest quarterly increase on record. At a time when people believed banks were contributing the

least to the economy since the 1930s, the National Accounts indicated the financial sector was contributing the most since the mid-1980s. How do we begin to square this circle?

That is the purpose of this chapter. It is planned as follows. In Section 2, we consider conventional measures of financial sector value added and how these have evolved over time. In Section 3, we consider a growth accounting breakdown of the factor inputs which have driven growth – quantities of labour and capital and the returns to these factors. This suggests banking has undergone, at least arithmetically, a "productivity miracle" over the past few decades. Section 4 explores in greater detail some of the quantitative drivers of high aggregate returns to banking, while Section 5 explores some of banks' business activities. Risk illusion, rather than a productivity miracle, appears to have driven high returns to finance. The recent history of banking appears to be as much mirage as miracle. Section 6 concludes with some policy implications.

2. Measuring financial sector output

(a) Historical trends in GVA

The standard way of measuring the contribution of a sector to output in the economy is GVA. This is defined as the value of gross output that a sector or industry produces less the value of intermediate consumption (that is, goods and services used in the process of production). GVA only measures the sector's direct contribution to the economy. The indirect contribution of finance – for example, on productivity growth through the provision of funds for start-up businesses and new investment projects – may also be important. But looking at historical trends in value added is a useful starting point.

Chart 1 plots an index of real GVA of the financial intermediation sector in the UK from the middle of the 19th century, alongside an index of whole-economy output. Both series are in constant prices and indexed to 1975=100. Table 1 breaks down the growth rates of finance and whole economy output into three sub-samples – pre-First World War, from the First World War to the early 1970s, and thence to date. The historical trends in GVA for the financial sector are striking.

Over the past 160 years, growth in financial intermediation has outstripped whole economy growth by over 2 percentage points per year. Or put differently, growth in financial sector value added has been more than double that of the economy as a whole since 1850. This is unsurprising in some respects. It reflects a trend towards financial deepening which is evident across most developed and developing economies over the past century. This structural trend in finance has been shown to have contributed positively to growth in the whole-economy (Wadhwani (2010)).

The sub-sample evidence suggests, however, that this has not been a straight line trend. The pre-First World War period marked a period of very rapid financial deepening, with the emergence of joint stock banks to service the needs of a rapidly growing non-financial economy. Finance grew at almost four times the pace of the real economy during this rapid-growth period (Table 1).

The period which followed, from the First World War right through until the start of the 1970s, reversed this trend. The growth in finance fell somewhat short of that in the rest of the economy. This in part reflected the effects of tight quantitative constraints on, and government regulation of, the financial sector.

The period from the early 1970s up until 2007 marked another watershed. Financial liberalisation took hold in successive waves. Since then, finance has comfortably outpaced growth in the non-financial economy, by around 1.5 percentage points per year. If anything, this trend accelerated from the early 1980s onwards. Measured real value added of the financial intermediation sector more than trebled between 1980 and 2008, while whole economy output doubled over the same period.

In 2007, financial intermediation accounted for more than 8% of total GVA, compared with 5% in 1970. The gross operating surpluses of financial intermediaries show an even more dramatic trend. Between 1948 and 1978, intermediation accounted on average for around 1.5% of whole economy profits. By 2008, that ratio had risen tenfold to about 15% (Chart 2).

Internationally, a broadly similar pattern is evident. In the US, following a major decline during the Great Depression, the value added of the financial sector has risen steadily since the end of the Second World War. As a fraction of whole economy GVA, it has quadrupled over the period, from about 2% of total GDP in the 1950s to about 8% today (Chart 3). Similar trends are evident in Europe and Asia. According to data from the *Banker*, the largest 1000 banks in the world reported aggregate pre-tax profits of almost $800 billion in fiscal year 2007/08 (Chart 4), almost 150% higher than in 2000/01. This equates to annualised returns to banking of almost 15%.

Some of these trends in the value added and profits of the financial sector, and in particular their explosive growth recently, are also discernible in the market valuations of financial firms relative to non-financial firms. Total returns to holders of major banks' equity in the UK, US and euro area rose a cumulative 150% between 2002 and 2007 (Chart 5). This comfortably exceeded the returns to the non-financial economy and even to some of the more risk-seeking parts of the financial sector, such as hedge funds.

To illustrate this rather starkly, consider a hedged bet placed back in 1900, which involved going long by £100 in financial sector equities and short in non-financial equities by the same amount. Chart 6 shows cumulative returns to following this hedged strategy. From 1900 up until the end of the 1970s, this bet yielded pretty much nothing, with financial and non-financial returns rising and falling roughly in lockstep. But from then until 2007, cumulative returns to finance took off and exploded in a bubble-like fashion. Only latterly, with the onset of the crisis, has that bubble burst and returned to earth.

(b) Measuring GVA in the financial sector

To begin to understand these trends, it is important first to assess how financial sector value-added is currently measured and the problems this poses when gauging the sector's contribution to the broader economy.

Most sectors charge explicitly for the products or services they provide and are charged explicitly for the inputs they purchase. This allows the value-added of each sector to be measured more or less directly. For example, gross output of a second-hand car dealer can be calculated as the cash value of all cars sold. The value added of that dealer would then be estimated by subtracting its intermediate consumption (the value of cars bought) from gross output.

This is also the case for some of the services provided by the financial sector.[1] For example, investment banks charge explicit fees when they advise clients on a merger or acquisition. Fees or commissions are also levied on underwriting the issuance of securities and for the market-making activities undertaken for clients. But such direct charges account for only part of the financial system's total revenues. Finance – and commercial banking in particular – relies heavily on interest flows as a means of payment for the services they provide. Banks charge an interest rate margin to capture these intermediation services.

To measure the value of financial services embedded in interest rate margins, the concept of FISIM – Financial Intermediation Services Indirectly Measured – has been developed internationally. The concept itself was introduced in the 1993 update of the United Nations System of National Accounts (SNA). The SNA recognises that financial intermediaries

[1] For further details refer to, for example, Akritidis L (2007).

provide services to consumers, businesses, governments and the rest of the world for which explicit charges are not made. In associated guidelines, a number of such services are identified including:

- Taking, managing and transferring deposits;
- Providing flexible payment mechanisms such as debit cards;
- Making loans or other investments; and
- Offering financial advice or other business services.

FISIM is estimated for loans and deposits only. The calculation is based on the difference between the effective rates of interest (payable and receivable) and a "reference" rate of interest, multiplied by the stock of outstanding balances. According to SNA guidelines, "this reference rate represents the pure cost of borrowing funds – that is, a rate from which the risk premium has been eliminated to the greatest extent possible, and that does not include any intermediation services".[2] For example, a £1,000 loan with a 9% interest receivable and a 4% reference rate gives current price FISIM on the loan = £1,000 x (9%–4%) = £50. And for a £1,000 deposit with a 3% interest payable and a 4% reference rate, this gives current price FISIM on the deposit = £1,000 x (4%–3%) = £10. Overall, estimated current price FISIM accounts for a significant share of gross output of the banking sector (Chart 7).

Estimating a real measure of FISIM is fraught with both conceptual and computational difficulties. In the earlier example of the second-hand car dealer, statisticians can use the number of cars sold as an indicator of the volume of gross output. But the conceptual equivalent for financial intermediation is not clear. Would two loans of £50 each to the same customer represent a higher level of activity than one loan of £100? Methods for measuring FISIM at constant prices are based on conventions. In the UK, real FISIM is calculated by applying the base-year interest margins to an appropriate volume indicator of loans and deposits. The latter is estimated by deflating the corresponding stocks of loans and deposits using the GDP deflator. This method means that any volatility in the current price measure of FISIM caused by changes in interest margins does not feed into the real measure.

(c) Refining the measurement of FISIM

While the introduction of FISIM into the national accounts was an important step forward, it is not difficult to construct scenarios where the contribution of the financial sector to the economy could be mis-measured under this approach. A key issue is the extent to which bearing risk should be measured as a productive service provided by the banking system.

(i) Adjusting FISIM for risk

Under current FISIM guidelines, which use risk-free policy rates to measure the reference rate, banks' compensation for bearing risk constitutes part of their measured nominal output.

---

[2] 1993 System of National Accounts, paragraph 6.128: http://unstats.un.org/unsd/sna1993/toctop2.asp.

This can lead to some surprising outcomes. For example, assume there is an economy-wide increase in the expected level of defaults on loans or in liquidity risk, as occurred in October 2008. Banks will rationally respond by increasing interest rates to cover the rise in expected losses. FISIM will score this increased compensation for expected losses on lending as a rise in output. In other words, at times when risk is rising, the contribution of the financial sector to the real economy may be overestimated. This goes some way towards explaining the 2008Q4 National Accounts paradox of a rapidly rising financial sector contribution to nominal GDP.

Of course, the financial sector does bear the risk of other agents in the economy. Banks take on maturity mismatch or liquidity risk on behalf of households and companies. And banks also make risky loans funded by debt, which exposes them to default or solvency risk. But it is not clear that bearing risk is, in itself, a productive activity. Any household or corporate investing in a risky debt security also bears credit and liquidity risk. The act of investing capital in a risky asset is a fundamental feature of capital markets and is not specific to the activities of banks. Conceptually, therefore, it is not clear that risk-based income flows should represent bank output.

The productive activity provided by an effectively functioning banking system might be better thought of as measuring and pricing credit and liquidity risk. For example, banks screen borrowers' creditworthiness when extending loans, thereby acting as delegated monitor. And they manage liquidity risk through their treasury operations, thereby acting as delegated treasurer. These risk-pricing services are remunerated implicitly through the interest rates banks charge to their customers.

Stripping out the compensation for bearing risk to better reflect the service component of the financial sector could be achieved in different ways. One possibility would be to adjust FISIM using provisions as an indicator of expected losses. A broader adjustment for risk, as has been suggested by several commentators, would be to move away from the risk-free rate as the reference rate within FISIM.[3] For example, a paper prepared for the OECD Working Party on National Accounts (Mink (2008)) suggested that the FISIM calculation should use reference rates that match the maturity and credit risk of loans and deposits. This would also eliminate an inconsistency within the current National Accounts framework. Measured financial intermediation output increases if a bank bears the risk of lending to a company. But gross output is unchanged if a household holds a bond issued by the same company and thus bears the same risk.

To see how such a mechanism would work, consider the following simple example. A bank lends £100 to a corporate borrower at 7% per annum for one-year. The risk-free rate is 5%. The bank correctly assesses the credit risk of the corporate to be A-rated. The market spread for A-rated credits at a maturity of one-year is 1% over the risk-free rate. Current FISIM would estimate bank output as £2 (Table 2). Risk-adjusted FISIM, though, would estimate banks' output as £1.

An adjustment of FISIM along these lines could potentially be material. According to simulations on the impact of such an approach for the Euro-area countries, aggregate risk-adjusted FISIM would stand at about 60% of current aggregate FISIM for the Euro-area countries over the period 2003–7 (Mink (2008)).

(ii) Measuring risk

Adjusting FISIM for risk would better capture the contribution of the financial sector to the economy. The fundamental problem is, however, that risk itself is unobservable ex-ante. The methodology described above measures risk in a relative way; it effectively assumes that if

[3] Wang et al (2004), Wang (2003), Mink (2008), Colangelo and Inklaar (2010).

banks deviate from prevailing market rates, this is to compensate for the services they provide to borrowers and depositors. But at no point is there an assessment of the ability of the financial system to price risk correctly in an absolute sense. This might not be the objective of statisticians when measuring output. But it is essential when gauging the contribution of finance to economic well-being.

To see this more clearly, consider an alternative example (Table 3). A bank lends £100 to a corporate borrower. But the bank incorrectly assesses the credit risk of the corporate to be A-rated, when the true credit risk is BB-rated. Assume for simplicity that the corporate, knowing that its credit risk is greater than A, is prepared to pay a spread higher than that on an A-rated credit risk (say 2%). The market spreads for A-rated and BB-rated credits are 1% and 2% respectively. "Measured" risk-adjusted FISIM is still an improvement on current FISIM. But the value of bank output is still overstated relative to "true" risk-adjusted FISIM.

This would be equivalent to second-car hand dealers consistently selling lemons. But a dodgy car-seller would be quickly found out. Mechanical risk is observable. Dealers that persistently mis-price cars would be driven from the market. Buyers might instead then choose to meet online.

A banking system that does not accurately assess and price risk is not adding much value to the economy. Buyers and sellers of risk could meet instead in capital markets – as they have, to some extent, following the crisis. But unlike the condition of a car, risk is unobservable. So mis-pricing of risk, and mis-measurement of the services banks provide to the real economy, may persist. This echoes events in the run-up to crisis when market prices systematically under-priced risk for a number of years. Using the market price of risk would have led statisticians systematically to overstate the potential contribution of the financial sector over this period.

Attempting to adjust the measurement of bank output for risk by changing the reference rate in FISIM is an improvement on current practices. But it would still fall short of assessing whether the financial sector is pricing risk correctly and hence assessing the true value of the services banks provides to the wider economy. Unless the price of risk can be evaluated, it seems unlikely the contribution of the financial sector to the economy can be measured with accuracy.

3. Decomposing the contribution of the financial sector – the productivity "miracle"

To that end, an alternative way of looking at the contribution of the financial sector is through inputs to the production process. This might shed more light on the sources of the rapid growth in finance. Was this expansion accompanied by a rising share of resources employed by finance relative to the rest of the economy? Or did it instead reflect unusually high returns to these factors of production? This section considers these questions in turn.

(a) Growth accounting decomposition

The basic growth accounting framework breaks down the sources of economic growth into the contributions from increases in the inputs to production, capital and labour. This amounts to relating growth in GDP to growth in labour input and in various capital services (from buildings, vehicles, computers and other resources). When these factors have all been accounted for, the remainder is often attributed to technical change – the so-called Solow residual (Solow (1957)).

The growth accounting framework assumes an underlying aggregate production function. In its most basic form, the aggregate production function can be written as:

$$Q = f(K, L, t)$$

where Q is output, K and L represent capital and labour units and t appears in *f* to allow for technical change.

Assuming constant returns to scale, perfect competition (so that factors of production are paid their marginal products) and Hicks-neutral technical change (so that shifts in the production function do not affect marginal rates of substitution between inputs), output growth can be expressed as a weighted sum of the growth rates of inputs and an additional term that captures shifts over time in the production technology. The weights for the input growth rates are the respective shares in total input payments – the labour and capital shares. More specifically:

$$\frac{\dot{Q}}{Q} = \frac{\dot{A}}{A} + \alpha_K \frac{\dot{K}}{K} + \alpha_L \frac{\dot{L}}{L}$$

where A(t) is a multiplicative factor in the production function capturing technical change. α_K, α_L represent respectively the capital and labour shares of income.

Charts 8 and 9 look at the proportion of labour and physical capital employed by the financial intermediation sector in the UK relative to the whole economy over the past forty years. They follow a not dissimilar path, with both labour and capital inputs rising as a share of the whole economy for much of the period. The proportion of labour employed by finance rises by around 50% between 1977 and 1990, while the proportion of capital almost trebles from 4% to 12% over the same period. Financial liberalisation over the period drew factors of production into finance, both labour and capital, on a fairly dramatic scale.

Perhaps the most striking development, however, is what happens next. These trends have not persisted during this century. If anything, the labour and capital shares of the financial sector have been on a gently declining path over this period. Growth in both labour and capital employed in the financial sector has been modest and has been lower than in the economy as a whole. Since this fall in factor input shares coincides with a period when measured value-added of the financial sector was rising sharply, this suggests something dramatic must have been happening to productivity in finance – the Solow residual.

The measured residual, in a growth accounting sense, reflects improvements in the total factor productivity (TFP) of the inputs. A growth accounting decomposition suggests that measured TFP growth in the financial sector averaged about 2.2% per year between 1995 and 2007 (Chart 10). This comfortably exceeds TFP growth at the whole-economy level, estimated at an average of about 0.5–1.0% over the same period. In other words, on the face of it at least, there is evidence of the financial sector having undergone something of a "productivity miracle" during this century. This pattern has not been specific to the UK. Measured TFP growth in the financial sector exceeded that of the whole economy across many developed countries between 1995–2007, a trend that accelerated in the "bubble" years of 2003–2007 (Chart 11).

(b) Returns to factors of production

TFP in a growth framework is no more than an accounting residual. It provides no explanation of the measured productivity "miracle" in finance. A related question is whether the observed productivity miracle was reflected in returns to the factors of production in finance. Chart 12 decomposes total GVA of financial corporations into income flowing to labour (defined to include employees only) and income flowing to capital. Broadly speaking, the rise in GVA is equally split between the returns to labour (employee compensation) and to capital (gross operating surplus). The miracle has been reflected in the returns to both labour and capital, if not in the quantities of these factors employed.

For labour, these high returns are evident both in cross-section and time-series data. Chart 13 shows average weekly earnings across a range of sectors in the UK in 2007.

Financial intermediation is at the top of the table, with weekly average earnings roughly double those of the whole-economy median. This differential widened during this century, broadly mirroring the accumulation of leverage within the financial sector (Chart 14).

The time-series evidence is in some respects even more dramatic. Philippon and Reshef (2009) have undertaken a careful study of "excess" wages in the US financial industry since the start of the previous century, relative to a benchmark wage. Chart 15 plots their measure of excess wages. This shows a dramatic spike upwards which commenced in the early 1980s, but which exploded from the 1990s onwards. The only equivalent wage spike was in the run-up to the Great Crash in 1929. Philippon and Reshef attribute both of these wage spikes to financial deregulation.

This picture is broadly mirrored when turning from returns to labour to returns to capital. In the 1950s gross profitability of the financial sector relative to capital employed was broadly in line with the rest of the economy (Chart 16). But since then, and in particular over the past decade, returns to capital have far outpaced those at an economy-wide level.

Chart 17 plots UK banks' return on equity capital (ROE) since 1920 (Alessandri and Haldane (2009)). Although conceptually a different measure of returns to capital, the broad message is the same. Trends in ROE are clearly divided into two periods. In the period up until around 1970, ROE in banking was around 7% with a low variance. In other words, returns to finance broadly mimicked those in the economy as whole, in line with the gamble payoffs in Chart 6. But the 1970s mark a regime shift, with the ROE in banking roughly trebling to over 20%, again in line with gamble payoffs. Excess returns accumulated to capital as well as labour.

These returns were by no means unique to UK banks. Chart 18 plots ROEs for major internationally active banks in the US and Europe during this century. Two features are striking. First, the level of ROEs was consistently at or above 20% and on a rising trend up until the crisis. This is roughly double ROEs in the non-financial sector over the period. Second, the degree of cross-country similarity in these ROE profiles is striking. This, too, is no coincidence. During much of this period, banks internationally were engaged in a highly competitive ROE race. Therein lies part of the explanation for these high returns to labour and capital in banking.

4. Explaining aggregate returns in banking – excess returns and risk illusion

How do we explain these high, but temporary, excess returns to finance which appear to have driven the growing contribution of the financial sector to aggregate economic activity? In this section we discuss potential balance sheet strategies which may have contributed to these rents. Essentially, high returns to finance may have been driven by banks assuming higher risk. Banks' profits, like their contribution to GDP, may have been flattered by the mis-measurement of risk.

The crisis has subsequently exposed the extent of this increased risk-taking by banks. In particular, three (often related) balance sheet strategies for boosting risks and returns to banking were dominant in the run-up to crisis:

- increased leverage, on and off-balance sheet;
- increased share of assets held at fair value; and
- writing deep out-of-the-money options.

What each of these strategies had in common was that they generated a rise in balance sheet risk, as well as return. As importantly, this increase in risk was to some extent hidden by the opacity of accounting disclosures or the complexity of the products involved. This resulted in a divergence between reported and risk-adjusted returns. In other words, while reported ROEs rose, *risk-adjusted* ROEs did not (Haldane (2009)).

To some extent, these strategies and their implications were captured to a degree in performance measures. For example, the rise in reported average ROEs of banks over the past few decades occurred alongside a rise in its variability. At the same time as average ROEs in banking were trebling, so too was their standard deviation (Chart 17). In that sense, the banking "productivity miracle" may have been, at least in part, a mirage – a simple, if dramatic, case of risk illusion by banks, investors and regulators.

(a) Increased leverage

Banks' balance sheets have grown dramatically in relation to underlying economic activity over the past century. Charts 19 and 20 plot this ratio for the UK and the US over the past 130 years. For the US, there has been a secular rise in banks' assets from around 20% to over 100% of GDP. For the UK, a century of flat-lining at around 50% of GDP was broken in the early 1970s, since when banks' assets in relation to national income have risen tenfold to over 500% of GDP.

This century has seen an intensification of this growth. According to data compiled by the *Banker*, the balance sheets of the world's largest 1000 banks increased by around 150% between 2001 and 2009 (Chart 21). In cross-section terms, the scale of assets in the banking system now dwarfs that in other sectors. Looking at the size of the largest firm's assets in relation to GDP across a spectrum of industries, finance is by far the largest (Chart 22).

The extent of balance sheet growth was, if anything, understated by banks' reported assets. Accounting and regulatory policies permitted banks to place certain exposures off-balance sheet, including special purpose vehicles and contingent credit commitments. Even disclosures of on-balance sheet positions on derivatives disguised some information about banks' contingent exposures.

This rapid expansion of the balance sheet of the banking system was not accompanied by a commensurate increase in its equity base. Over the same 130 year period, the capital ratios of banks in the US and UK fell from around 15–25% at the start of the 20th century to around 5% at its end (Chart 23). In other words, on this metric measures of balance sheet leverage rose from around 4-times equity capital in the early part of the previous century to around 20 times capital at the end.

If anything, the pressure to raise leverage increased further moving into this century. Measures of gearing rose sharply between 2000 and 2008 among the major global banks, other than US commercial banks which were subject to a leverage ratio constraint (Chart 24). Once adjustments are made to on- and off-balance sheet assets and capital to give a more comprehensive cross-country picture, levels of gearing are even more striking. Among the major global banks in the world, levels of leverage were on average more than 50 times equity at the peak of the boom (Chart 25).

For a given return on assets (RoA), higher leverage mechanically boosts a banks' ROE. The decision by many banks to increase leverage appears to have been driven, at least in part, by a desire to maintain ROE relative to competitors, even as RoA fell. For example, as Chart 26 illustrates, virtually all of the increase in the ROE of the major UK banks during this century appears to have been the result of higher leverage. Banks' return on assets – a more precise measure of their productivity – was flat or even falling over this period.

Between 1997 and 2008, as UK banks increased leverage, they managed to maintain broadly constant capital ratios by, on average, seeking out assets with lower risk weights (Chart 27). A similar pattern was evident among a number of the Continental European major global banks (Chart 28). It is possible to further decompose ROE to provide additional insight into how banks increased reported returns as follows:

$$RoE = \frac{Total\ assets}{Tier\ 1\ capital} X \frac{Tier\ 1\ capital}{Common\ equity} x \frac{Net\ income}{RWAs} x \frac{RWAs}{Total\ assets} \quad (1.1)$$

$$RoE = Financial\ leverage\ X\ Common\ equity\ margin\ x\ RoRWAs\ x\ Unit\text{-}risk$$

Banks can boost ROE by acting on any of the terms on the right-hand side of equation (1.1): increasing assets relative to capital (financial leverage), holding a larger proportion of capital[4] other than as common equity (common equity margin), or assuming a greater degree of risk per unit of assets (return on risk-weighted assets, RoRWA) – *leveraging assets, leveraging* capital structure or *leveraging* regulation.

Table 4 shows two of the elements of this breakdown for the major global banks – leverage and unit risk. For most banks, the story is one of a significant increase in assets relative to capital, with little movement into higher risk assets (unit risk makes a negative contribution for most banks). Those banks with highest leverage, however, are also the ones which have subsequently reported the largest write-downs. That suggests banks may also have invested in riskier assets, which regulatory risk-weights had failed to capture.

Table 5 looks at the third component, the common equity margin, of some of the same global banks. Among at least some of these banks, this margin makes a significant contribution to ROE growth, as banks moved into hybrid Tier 1 capital instruments at the expense of core equity. As such hybrid instruments have shown themselves largely unable to absorb losses during the crisis, this boost to ROE is also likely to have been an act of risk illusion.

Taken together, this evidence suggests that much of the "productivity miracle" of high ROEs in banking appear to have been the result not of productivity gains on the underlying asset pool, but rather a simple leveraging up of the underlying equity in the business.

(b) Larger trading books

A second strategy pursued by a number of banks in the run-up to crisis was to increase their assets held at fair value, principally through their trading books, relative to their banking books of underlying loans. Among the major global banks, the share of loans to customers in total assets fell from around 35% in 2000 to 29% by 2007 (Chart 29). Over the same period, trading book asset shares almost doubled from 20% to almost 40%. These large trading books were associated with high leverage among the world's largest banks (Chart 30).

What explains this shift in portfolio shares? Regulatory arbitrage appears to have been a significant factor. Trading book assets tended to attract risk weights appropriate for dealing with market but not credit risk. This meant it was capital-efficient for banks to bundle loans into tradable structured credit products for onward sale. Indeed, by securitising assets in this way, it was hypothetically possible for two banks to swap their underlying claims but for both firms to claim capital relief. The system as a whole would then be left holding less capital, even though its underlying exposures were identical. When the crisis came, tellingly losses on structured products were substantial (Chart 31).

A further amplifying factor is that trading books are marked-to-market and any gains or losses taken through to the profit and loss account. So holding a large trading book is a very good strategy when underlying asset prices in the economy are rising rapidly. This was precisely the set of the circumstances facing banks in the run-up to crisis, with asset prices driven higher by a search for yield among investors. In effect, this rising tide of asset price

[4] The term "Tier 1 capital" refers to the component of banks' regulatory capital comprising common equity and capital instruments close to common equity ("hybrid Tier 1 capital"), as defined by rules set out by regulators. For a discussion of the composition of UK banks' regulatory capital see Bank of England (2009a).

rises was booked as marked-to-market profits by banks holding assets in their trading book. Everyone, it appeared, was a winner.

But because these gains were driven by a mis-pricing of risk in the economy at large, trading book profits were in fact largely illusory. Once asset prices went into reverse during 2008 as risk was re-priced, trading book losses quickly materialised. Write-downs on structured products totalled $210 billion among the major global banks in 2008 alone.

(c) Writing deep out-of-the-money options

A third strategy, which boosted returns by silently assuming risk, arises from offering tail risk insurance. Banks can in a variety of ways assume tail risk on particular instruments – for example, by investing in high-default loan portfolios, the senior tranches of structured products or writing insurance through credit default swap (CDS) contracts. In each of these cases, the investor earns an above-normal yield or premium from assuming the risk. For as long as the risk does not materialise, returns can look riskless – a case of apparent "alpha". Until, that is, tail risk manifests itself, at which point losses can be very large.

There are many examples of banks pursuing essentially these strategies in the run-up to crisis. For example, investing in senior tranches of sub-prime loan securitisations is, in effect, equivalent to writing deep-out-of-the-money options, with high returns except in those tail states of the world when borrowers default en masse. It is unsurprising that issuance of asset-backed securities, including sub-prime RMBS (residential mortgage-backed securities), grew dramatically during the course of this century, easily outpacing Moore's Law (the benchmark for the growth in computing power since the invention of the transistor) (Chart 32).[5]

Tranched structured products, such as CDOs (collateralised debt obligations) and CLOs (collateralised loan obligations), generate a similar payoff profile for investors to sub-prime loans, yielding a positive return in stable states of the world – apparent alpha – and a large negative return in adverse states. Volumes outstanding of CDOs and CLOs also grew at a rate in excess of Moore's Law for much of this century. The resulting systematic mis-pricing of, in particular, the super-senior tranches of these securities was a significant source of losses to banks during the crisis, with ratings downgrades large and frequent (Chart 33).

A similar risk-taking strategy was the writing of explicit insurance contracts against such tail risks, for example through CDS. These too grew very rapidly ahead of crisis (Chart 34). Again, the writers of these insurance contracts gathered a steady source of premium income during the good times – apparently "excess returns". But this was typically more than offset by losses once bad states materialised. This, famously, was the strategy pursued by some of the monoline insurers and by AIG. For example, AIG's capital market business, which included its ill-fated financial products division, reported total operating income of $2.3 billion in the run-up to crisis from 2003 to 2006, but reported operating losses of around $40 billion in 2008 alone.

What all of these strategies had in common was that they involved banks assuming risk in the hunt for yield – risk that was often disguised because it was parked in the tail of the return distribution. Excess returns – from leverage, trading books and out-of-the-money options – were built on an inability to measure and price risk. The productivity miracle was in fact a risk illusion. In that respect, mis-measurement of the contribution of banking in the National Accounts and the mis-measurement of returns to banking in their own accounts have a common underlying cause.

---

[5] Moore's Law refers to the observation by Intel co-founder Gordon Moore in 1965 that transistor density on integrated circuits had doubled every year since the integrated circuit was invented and the prediction that this would continue.

5. Explaining disaggregated returns to banking

A distinct, but complementary, explanation of high returns to banking is that they reflect structural features of the financial sector. For example, measures of market concentration are often used as a proxy for the degree of market power producers have over consumers. It is telling that measures of the concentration of the banking sector have increased dramatically over the course of the past decade, coincident with the rise in banking returns. Chart 35 plots the share of total bank assets of the largest three banks in the US since the 1930s. Having flat-lined up until the 1990s, the top 3 share has since roughly tripled. A similar trend is evident in the UK (where the share of the top 3 banks currently stands at above 50%) and globally (where the share of the top 3 has doubled over the past 10 years).

At the same time, it is well known that market concentration need not signal a lack of competitiveness or efficiency within an industry or sector (Wood and Kabiri (2010)). Highly competitive industries can be concentrated and highly decentralised industries uncompetitive. A better arbiter of market power may be measures of market contestability, in particular the potential for barriers to entry to and exit from the market. Entry and exit rates from banking have, historically, tended to be very modest by comparison with the non-financial sector and other parts of the financial sector, such as hedge funds.

For banks operating in many markets and offering a range of services, aggregate returns may offer a misleading guide to the degree of market contestability. Looking separately at the different activities financial firms undertake provides a potentially clearer indication of the drivers of performance and the structural factors determining them. In this respect, JP Morgan Chase provides an interesting case study.

JP Morgan Chase is a large universal bank offering a full package of banking services to customers, retail and wholesale. Its published accounts also provide a fairly detailed decomposition of the returns to these different activities. Chart 36 looks at the returns on equity at JP Morgan Chase, broken down by business line and over time. These estimates are based on the firm's economic capital model. So provided this model adequately captures risk, these estimates ought to risk-adjust returns across the different business lines, allocating greater amounts of capital to riskier activities.

(a) "Low risk/low return" business activities

Consider first some of the activities generally perceived to be low-risk/low return – asset management and treasury and securities services and retail financial services. All of these seemingly low risk activities appear to deliver above-average returns on equity, ranging from a high of around 50% on treasury and asset management services to around 20%+ on retail financial services.

One potential explanation of these high returns is that the risk associated with these activities, and hence the capital allocated to them, may be under-estimated by banks' models. Another is that the demand for these services is highly price inelastic – for example, because of information imperfections on the part of end-users of these services. Anecdotally, there is certainly evidence of a high degree of stickiness in the demand for retail financial services. Statistically, an adult is more likely to leave their spouse than their bank.

In a UK context, there have been a number of studies by the authorities on the degree of competition within retail financial services, including by the Competition Commission (2005) and the Office of Fair Trading (OFT) (2008). The OFT market study found a very low rate of switching of personal current accounts between banks – fewer than 6% per year. By itself, however, this low switching rate does not necessarily imply a market failure. For example, it could be the result of a reputational equilibrium in which money gravitates to banks whose brand name is recognised and respected.

A more obvious market friction in the UK retail financial services market derives from "free in credit" banking. In effect, all retail payment services are charged at a zero up-front fee,

except large-value payment transfers through CHAPS[6] (which are typically charged at around £25). This charging schedule is not well aligned with marginal costs. It encourages bundling of payment services and the charging of latent or hidden fees on other transactions services – for example, overdraft fees. Explicit charging for retail financial services would increase transparency and reduce the scope for distortions in the use of these services.

High returns on treasury management services also present something of a puzzle. These include transactions, information and custodial services to clients. None of these activities are especially expertise-intensive and the market for these services ought in principle to be contestable internationally.

(b) "High risk/high return" business activities

The higher risk activities associated with finance, such as commercial and investment banking, do not on the face of it appear to yield as high returns on equity. Nonetheless these returns, at around 20%, are above levels in the non-financial sector.

Investment banking activities are, in risk terms, a mixed bag. They comprise fairly low-risk activities, such as (merger and acquisition) M&A advisory work, with higher-risk activities such as securities underwriting and proprietary trading. To complicate matters, banks' annual accounts data do not differentiate simply between these activities – for example, between market-making and proprietary trading activities in fixed income, currency and commodities (FICC) and equities. Chart 37 provides a revenue breakdown of US investment banks' activities.

The lack of a breakdown between client and proprietary sources of revenues is problematic when making sense of investment banking activities, both in the run-up to and during the crisis. In the run-up to crisis, FICC and equity-related activity contributed significantly to revenues, partly on the back of proprietary trading in assets whose prices were rising rapidly. Some of these gains then dissolved when asset prices, in particular for FICC, went into reverse during 2008.

The story of 2009/10 is of a strong recovery in FICC and equity revenues. The source of this revenue recovery is, however, different to the boom. Instead of proprietary risk-taking, increased revenues appear instead to have been driven by market-making activities on behalf of clients. These were boosted by a bulge in client activity and wider bid-ask spreads, against a backdrop of lower levels of competition (Chart 38). It is an open question whether these returns to market-making will persist.

In some respects, returns to M&A and advisory activities represent even more of a puzzle. For a start, it is well known that most M&A activity is value-destroying (for example, Palia (1995)). Advisory fees of 0.5–1.5% are typically taken, even though these activities are essentially risk-less. And in total under-writing fees are often around 3–4% in Europe and higher still in the US, having risen during the course of the crisis. The level and persistence of these fees is also something of a puzzle.

One potential explanation is that high fees on underwriting and advisory activities are sustained as a reputational equilibrium. In effect, clients are willing to pay a premium to have bonds or equity underwritten by a recognised name, as this is a signal of quality to end-investors. A similar phenomenon might explain the "2 and 20" fee structure of hedge funds. The OFT has recently announced an investigation into underwriting fees in the UK market.

---

[6] CHAPS is the same-day electronic funds transfer system, operated by the bank-owned CHAPS Clearing Company, that is used for high-value/wholesale payments but also for other time-critical lower value payments (such as house purchase).

Another part of the puzzle was banks' approach to managing risk across these business lines. For example, treasury functions are designed to help a firm as a whole manage its balance sheet, with internal transfer pricing for liquidity services to business lines. By acting in that way, the risk-taking incentives of each business unit can be aligned with the business as a whole, thereby complementing firms' internal risk management.

In practice, during the run-up to crisis, treasury functions were often run as a profit centre. That would tend to encourage two sets of risk-taking behaviour. First, it may have encouraged banks to take risks in balance sheet management – for example, by seeking out cheaper sources of capital (for example, hybrids over pure equity) or liquidity (shorter-term unsecured borrowing over long-term secured funding). Second, it may have led to the systematic under-pricing of liquidity services to banks' business unit, fuelling excessive growth and/or risk-taking. Tackling these risks would require banks' treasury operations to cease being profit centres and to execute effective internal transfer pricing.

6. Conclusion

The financial sector has undergone an astonishing roller-coaster in the course of a decade. The ascent to heaven and subsequent descent to hell has been every bit as dramatic as in the 1930s. In seeking to smooth next time's ride, prophylactic public policy has a key role to play. Of the many initiatives that are underway, this paper has highlighted three which may warrant further attention in the period head:

- First, given its ability to both invigorate and incapacitate large parts of the non-financial economy, there is a strong case for seeking improved means of measuring the true value-added by the financial sector. As it is rudimentary to its activities, finding a more sophisticated approach to measuring risk, as well as return, within the financial sector would seem to be a priority. The conflation of the two can lead to an overstatement of banks' contribution to the economy and an understatement of the true risk facing banks and the economy at large. Better aggregate statistics and bank-specific performance measures could help better to distinguish miracles and mirages. This might include developing more sophisticated risk-adjustments to FISIM and a greater focus on banks' return on assets rather than equity by investors and managers.
- Second, because banks are in the risk business it should be no surprise that the run-up to crisis was hallmarked by imaginative ways of manufacturing this commodity, with a view to boosting returns to labour and capital. Risk illusion is no accident; it is there by design. It is in bank managers' interest to make mirages seem like miracles. Regulatory measures are being put in place to block off last time's risk strategies, including through re-calibrated leverage and capital ratios. But risk migrates to where regulation is weakest, so there are natural limits to what regulatory strategies can reasonably achieve. At the height of a boom, both regulators and the regulated are prone to believe in miracles. That is why the debate about potential structural reform of finance is important – to lessen the burden on regulation and reverse its descent into ever-greater intrusiveness and complexity. At the same time, regulators need also to be mindful of risk migrating outside the perimeter of regulation, where it will almost certainly not be measured.
- Third, finance is anything but monolithic. But understanding of these different business lines is complicated by the absence of reliable data on many of these activities. There are several open questions about the some of these activities, not least those for which returns appear to be high. This includes questions about the risks they embody and about the competitive structure of the markets in which they are traded. These are issues for both prudential regulators and the competition authorities, working in tandem. If experience after the Great Depression is any

guide, it seems likely that these structural issues will take centre-stage in the period ahead.

References

Akritidis, L (2007), "Improving the measurement of banking services in the UK National Accounts", *Economic and Labour Market Review* 1(5), pp. 29–37.

Alessandri, P and Haldane, A G (2009), *Banking on the State*, available at http://www.bankofengland.co.uk/publications/speeches/2009/speech409.pdf

Bank of England (2009a), "The changing composition of the major UK banks' regulatory capital", *Bank of England Financial Stability Report*, June, pp. 26–27.

Bank of England (2009b), *The Role of Macroprudential Policy – A Discussion Paper*, available at http://www.bankofengland.co.uk/publications/other/financialstability/roleofmacroprudentialpolicy091121.pdf

Berger, A, Herring, R and Szegö, G (1995), The Role of Capital in Financial Institutions, *Journal of Banking and Finance* Vol. 19(3–4), pp. 393–430.

Billings, M and Capie, F (2004), "Evidence on competition in English commercial banking, 1920–1970", *Financial History Review* Vol. 11.

Billings, M and Capie, F (2007), "Capital in British banking, 1920–1970", *Business History*, Vol. 49(2), pp. 139–162.

Colangelo, A and Inklaar, R (2010), "Banking Sector Output Measurement in the Euro Area – A Modified Approach", *ECB Working Paper Series* No. 1204, available at http://www.ecb.int/pub/pdf/scpwps/ecbwp1204.pdf

Competition Commission (2005), *Store Cards Market Inquiry: Provisional Findings Report*, available at http://www.competition-commission.org.uk/inquiries/completed/2006/storecard/provisional_findings.htm

Feinstein, C H (1972), *National Income, Expenditure and Output of the United Kingdom 1855–1965*, Cambridge University Press.

Goodhart, C (2010), "How should we regulate the financial sector?", *Future of Finance and the Theory That Underpins It.*

Haldane, A G (2009), *Small Lessons from a Big Crisis,* available at http://www.bankofengland.co.uk/publications/speeches/2009/speech397.pdf

Haldane, A G (2010), *The $100 Billion Question*, available at http://www.bankofengland.co.uk/publications/speeches/2010/speech433.pdf

Kay, J (2009), *Narrow Banking: The Reform of Banking Regulation,* Centre for the Study of Financial Innovation.

Large, A (2010), *Systemic Policy and Financial Stability: A Framework for Delivery*, Centre for the Study of Financial Innovation.

Mink, R (2008), *An Enhanced Methodology of Compiling Financial Intermediation Services Indirectly Measured (FISIM)*, paper presented at OECD Working Party on National Accounts, Paris, 14–16 October 2008, available at http://www.olis.oecd.org/olis/2008doc.nsf/LinkTo/NT000059AE/$FILE/JT03251258.PDF

Mitchell, B R (1988), *British Historical Statistics*, Cambridge University Press.

Office of Fair Trading (2008), *Personal Current Accounts in the UK – An OFT Market Study*, available at http://www.oft.gov.uk/shared_oft/reports/financial_products/OFT1005.pdf

O'Mahony, M and Marcel P T (2009), "Output, Input and Productivity Measures at the Industry Level: the EU KLEMS Database", *Economic Journal* 119(538), pp. 374–403.

Oulton, N and Srinivasan, S (2005), "Productivity growth in UK industries, 1970–2000: structural change and the role of ICT", *Bank of England Working Paper Series* No. 259.

Philippon, T (2008), "The Evolution of the US Financial Industry from 1860 to 2007: Theory and Evidence", available at http://pages.stern.nyu.edu/~tphilipp/papers/finsize.pdf.

Philippon, T and Reshef, A (2009), "Wages and Human Capital in the U.S. Financial Industry: 1909–2006", *NBER Working Paper Series* No. 14644.

Reinhart, C M and Rogoff, K (2009), *This Time is Different: Eight Centuries of Financial Folly*, Princeton University Press.

Schularick M and Taylor A M (2009), "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles and Financial Crises, 1870–2008", *NBER Working Paper Series* No. 15512.

Sheppard, D K (1971), *The Growth and Role of U.K. Financial Institutions 1880–1962*, Methuen.

Solow, R M (1957), "Technical Change and the Aggregate Production Function", *Review of Economics and Statistics* 39(3), pp. 312–320.

Wadhwani, S (2010), "What mix of monetary policy and regulation is best for stabilising the economy?", *Future of Finance and the Theory That Underpins It.*

Wang, J C (2003), "Loanable Funds, Risk, and Bank Service Output", *Federal Reserve Bank of Boston Working Paper Series* No. 03–4.

Wang, J C, Basu, A and Fernald J G (2004), *A General-Equilibrium Asset-Pricing Approach to the Measurement of Nominal and Real Bank Output*, Invited for conference volume on Price Index Concepts and Measurement, Conference on Research on Income and Wealth (CRIW), available at http://www.bos.frb.org/economic/wp/wp2004/wp047.htm

Wood, G and Kabiri, A (2010), *Firm Stability and System Stability: The Regulatory Delusion*, Paper prepared for a conference on Managing Systemic Risk at the University of Warwick 7th–9th April 2010.