Nonneutrality of Money in Classical Monetary Thought

Thomas M. Humphrey

Introduction

The rise of the new classical macroeconomics, with its key idea that systematic monetary policy cannot influence real activity, has revived interest in the so-called classical neutrality postulate. That postulate, of course, holds that money-stock changes affect only the price level and not real output and employment. My concern in this paper is not with the neutrality postulate per se but rather with some recent claims made about the original classical economists' adherence to it.

In particular, I am concerned with the contention that the classicals—i.e., those predominantly British economists who wrote during the period 1750-1870 dating roughly from the publication of David Hume's *Essays* to the emergence of the marginalist revolution in the writings of William Stanley Jevons, Carl Menger, and Leon Walras—denied that money-stock changes had output and employment effects even in the short run. Such contentions have been voiced most recently by Lucas Papademos and Franco Modigliani in their essay "The Supply of Money and the Control of Nominal Income" in volume 1 of the prestigious *Handbook of Monetary Economics*. They state:

> The role of money in classical economics is a simple one, and so is the effect of a change in the quantity of money on aggregate nominal income. According to classical theory all markets for goods, including the market for labour services, clear continuously, with relative prices adjusting flexibly to ensure the attainment of equilibrium. Resources are fully utilized and thus aggregate employment and output are always at the "full-employment" or "natural" levels determined by tastes, productive technology and endowments, except for transitory deviations due to real disturbances.
>
> In such an economy, money . . . does not influence the determination of relative prices, real interest rates, the equilibrium quantities of commodities, and thus aggregate real income. Money is "neutral", a "veil" with no consequences for real economic magnitudes . . . (pp. 405-6).

Others arguing that the classicals believed that money is always neutral with respect to output and employment include David Glasner, Arjo Klamer, Kevin Hoover, and Michael Artis. Glasner, in his 1989 book *Free Banking and Monetary Reform*, asserts that "in the economy the classical theorists envisioned, the monetary sector could not . . . be a source of instability. A disturbance could only arise in the nonmonetary or real sector . . ." (p. 59). Arjo Klamer agrees. In the first chapter of his well-known 1984 *Conversations with Economists*, he characterizes the classical view by means of a vertical aggregate supply schedule drawn at the full-capacity level of output in price-output space. The vertical supply curve guarantees that any money-induced shift in aggregate demand affects only the price level but not real output. Support for Klamer's interpretation comes from Kevin Hoover who, in his 1988 *The New Classical Macroeconomics: A Skeptical Enquiry*, writes:

> The vertical aggregate supply curve provides an adequate capsulization of the classical view. . . . Changes in the level of the stock of money would change the general level of prices, but, because money was thought to be neutral . . . relative prices and the levels of employment and output would not be affected (pp. 9-10).

Likewise, Michael Artis, in his 1984 *Macroeconomics*, explains:

> the classical model guarantees full employment equilibrium, and the 'neutrality of money', i.e. the property that changes in the nominal money supply do not affect the real outcomes, but only the price level (p. 193).

This article argues (1) that the foregoing interpretations are wrong, (2) that the classicals held that money affects output and employment certainly in the short run and perhaps to some extent in the long run too, (3) that they identified at least nine reasons for the occurrence of such effects, and (4) that their concern with money's impact on the level of real activity strongly influenced their views of the desirability or undesirability of monetary expansion and contraction. In short, the following survey of eleven leading classical monetary theorists—including Thomas Attwood, Jeremy Bentham, David Hume, Thomas Robert Malthus, John Ramsay McCulloch, James Mill, John Stuart Mill, David Ricardo, Henry Thornton, Robert Torrens, and John Wheatley—

reveals that at least eight rejected the notion that money is always neutral and that continuous market-clearing and perfect wage-price flexibility prevail.[1] In holding that money's short-run impact is predominantly on output while its long-run impact is chiefly on prices, the classicals adhered to much the same view expressed by Milton Friedman in his 1970 Wincott Memorial lecture on *The Counter-Revolution in Monetary Theory*. Wrote Friedman: "In the short run, which may be as much as five or ten years, monetary changes affect primarily output. Over decades, on the other hand, the rate of monetary growth affects primarily prices" (pp. 23-24).

The article proceeds as follows: First it itemizes the particular sources or causes of nonneutrality specified by the classicals. Next it describes what individual classical writers had to say about each item. Finally it shows how classical views of nonneutrality continue to survive in twentieth-century monetary thought. The central message is that the notion of at least some nonneutrality is part of an enduring classical monetary tradition and that theories stressing neutrality-always are a departure from that tradition.

Sources of Nonneutrality

The table below lists the causes of nonneutrality specified by the classicals. A glance at the table shows how erroneous is the notion that those economists denied that money affects real activity. For example, they argued that real effects could stem from price inertia which caused money-stock changes to influence output before fully affecting prices. They found another source of nonneutrality in the lag of nominal wages behind rising or falling prices. This lag caused real wages and thus real profits to change, thereby altering incentives for employment and production. They also attributed money's nonneutrality to the fixity of certain nominal contractual costs whose real burden rose or fell with deflation or inflation.

Inflation-induced shifts of real income from workers and rentiers to producers who invest in real capital constituted an additional source of nonneutrality. So did the lag in nominal interest rates behind inflation which caused real rates to change thus affecting business borrowing, capital investment, and real activity. Nonneutrality was also seen to stem from desired fixed inventory-to-sales ratios that transformed money-induced increases in sales into increased production for inventory. The classicals likewise traced nonneutrality to a confusion between changes in general and relative prices—this confusion causing monetary shocks to be misperceived as real ones requiring output adjustments.

The classicals further argued that money affects output by influencing business confidence. They also cited the boost to productivity given by money-induced increases in aggregate demand which, by extending the scope of the market for goods, encourages specialization and division of labor. Some classicals even held that money's output effects emanate from the need to work harder to maintain one's real income in the face of inflation.

Rightly or wrongly, the classicals appealed to many explanations to account for money's impact on

SOURCES OF NONNEUTRALITY

Source	Cause(s) Money to affect real activity through:	Described by:
Sticky prices	real expenditure	Hume
Sticky nominal wages	real wages	Thornton, Torrens
Fixed nominal costs	real cost burdens	Attwood, McCulloch
Fixed nominal income of certain groups ("forced saving")	distributive shares and capital formation	Bentham, Thornton, Malthus, Ricardo, McCulloch
Sticky nominal interest rates	real interest rates	Torrens
Fixed inventory-to-sales ratios	inventory investment	Thornton
General price-relative price confusion	misperceived price signals	J. S. Mill
State of business confidence	changes in confidence	Attwood, McCulloch, Torrens
Market-size limitation to division of labor	labor productivity	Attwood, Malthus, Torrens
Efforts to maintain real income	labor-force participation rate	Torrens

[1] On these points see O'Brien (1975, pp. 162-65) and Niehans (1987) both of whom stress the short-run nonneutrality of money in classical thought. See also Viner (1937, pp. 185-200) for an earlier treatment of that same subject.

output and employment. One of the first to do so was David Hume, who invoked the notion of price inertia.

David Hume and the Lag of Prices Behind Money

The classical theory of nonneutrality, though partly rooted in the writings of Richard Cantillon, John Law, and William Potter, owes its greatest debt to David Hume. In his 1752 essays "Of Money" and "Of Interest," Hume argued that while a fixed absolute quantity of money is of no consequence for the level of output and employment, *changes* in the quantity of money have a very real significance.

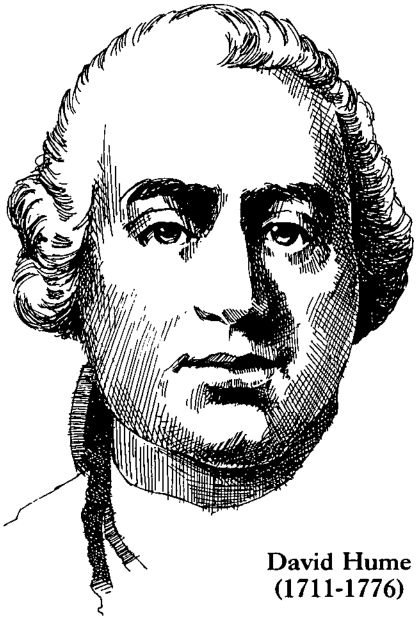
> Accordingly we find, that, in every kingdom into which money begins to flow in greater abundance than formerly, every thing takes a new face: labour and industry gain life; the merchant becomes more enterprising, the manufacturer more diligent and skilful, and even the farmer follows his plough with greater alacrity and attention (p. 37).

Hume attributes these nonneutralities to the lag of prices behind money. This lag, he says, causes money-induced changes in nominal spending to be divided in favor of output before being fully absorbed by prices. In his words:

> To account, then, for this phenomenon, we must consider, that though the high price of commodities be a necessary consequence of the encrease of gold and silver, yet it follows not immediately upon that encrease; but some time is required before the money circulates through the whole state, and makes its effect be felt on all ranks of people. At first, no alteration is perceived; by degrees the price rises, first of one commodity, then of another; till the whole at last reaches a just proportion with the new quantity of specie which is in the kingdom. In my opinion, it is only in this interval or intermediate situation, between the acquisition of money and rise of prices, that the encreasing quantity of gold and silver is favourable to industry (pp. 37-38).

Hume ascribes the price lag to the availability of idle labor willing to work at existing wages. Prices and wages rise only after all hands become fully employed.

> When any quantity of money is imported into a nation, it is not at first dispersed into many hands, but is confined to the coffers of a few persons, who immediately seek to employ it to advantage. . . . They are thereby enabled to employ more workmen than formerly, who never dream of demanding higher wages, but are glad of employment from such good paymasters. If workmen become scarce, the manufacturer gives higher wages, but at first requires an encrease of labour; and this is willingly submitted to by the artisan, who can now eat and drink better, to compensate his additional toil and fatigue. He carries his money to market, where he finds every thing at the same price as formerly, but returns with greater quantity and of better kinds, for the use of his family. . . . It is easy to trace the money in its progress through the whole commonwealth; where we shall find, that it must first quicken the diligence of every individual, before it encrease the price of labour (p. 38).



David Hume
(1711-1776)

Hume next distinguishes between temporary and permanent nonneutrality. Temporary nonneutrality stems from one-time changes in the money stock, changes to which prices eventually adjust. By contrast, permanent nonneutrality stems from a continuous succession of such changes to which prices never fully catch up.

As an example of temporary nonneutrality, Hume considers the transitory stimulus to output exerted by a one-time rise in the money stock. Noting that the stimulus vanishes once prices adjust to the augmented quantity of money, he concludes that

> Money, however plentiful, has no other effect, *if fixed*, than to raise the price of labour. . . . and . . . commodities. . . . In the progress towards these changes, the augmentation may have some influence, by exciting industry; but after the prices are settled, suitably to the new abundance of gold and silver, it has no manner of influence (pp. 47-48).

Hume points out that this same process works in reverse, a one-time contraction in the money stock first depressing output and employment before it lowers prices.

> A nation, whose money decreases, is actually, at that time, weaker and more miserable than another nation, which possesses no more money, but is on the encreasing hand. This will be easily accounted for, if we consider, that the alterations in the quantity of money . . . are not immediately attended with proportionable alterations in the price of

> commodities. There is always an interval before matters be adjusted to their new situation; and this interval is as pernicious to industry, when gold and silver are diminishing, as it is advantageous when these metals are encreasing (p. 40).

To Hume, monetary contraction had devastating real effects:

> The workman has not the same employment from the manufacturer and merchant; though he pays the same price for everything in the market. The farmer cannot dispose of his corn and cattle; though he must pay the same rent to his landlord. The poverty, and beggary, and sloth, which must ensue, are easily foreseen (p. 40).

Here is the source of the classicals' emphasis on the evils of monetary contraction.

As for permanent nonneutrality associated with sustained rates of monetary change, Hume argued as follows: Continuous money growth combines with sluggish price adjustment to keep money forever marching a step ahead of prices, perpetually frustrating the latter's attempts to catch up. The gap between money and prices persists indefinitely, thus producing a permanent change in the level of real activity. Hume's advice to the policymakers: exploit such nonneutrality via gradual enduring monetary expansion. For while

> it is of no manner of consequence, with regard to the domestic happiness of a state, whether money be in a greater or less quantity, [t]he good policy of the magistrate consists only in keeping it, if possible, still encreasing; because, by that means, he keeps alive a spirit of industry in the nation, and encreases the stock of labour, in which consists all real power and riches (pp. 39-40).

Hume's theory of the inflation mechanism was inherited by the other classical economists. Of these, only James Mill, David Ricardo, and John Wheatley rejected it in its entirety. Ricardo, whose skepticism of monetary policy's ability to influence real activity rivals that of modern new classicals, simply called Hume's theory "an erroneous view" (*Works*, V, 524) and remarked that "money cannot call forth goods" (*Works*, III, 301). Mill likewise dismissed Hume's mechanism with the assertion that money cannot exert even the briefest stimulus to output since prices instantly rise to absorb all the stimulus.[2] Wheatley was equally adamant, holding that "an increase of money has no other effect than to cause its own depression" in value (1803, p. 17, as quoted in Fetter 1942, p. 370).

True, Ricardo and Wheatley sometimes expressed concern with the evils of monetary contraction. But the evils they had in mind consisted almost solely of the arbitrary redistributive effects of deflation. Virtually no output or employment effects were envisioned.[3] Such views, however, were exceptions and not at all representative of the dominant classical position. Starting with Hume, most classicals accepted the view that money matters for real output and employment, temporarily if not permanently.

Lag of Wages Behind Prices

Hume blamed nonneutrality on sluggish price adjustment. The next source of nonneutrality recognized by the classicals was the lag of nominal wages behind prices. The classicals explained how monetary expansion and the resulting rise of prices would, because of the stickiness of wages relative to prices, lower real wages, raise real profits, and thereby spur

David Ricardo
(1772-1823)

[2] Mill wrote: "The man who goes first to market with the augmented quantity of money, either raises the price of the commodities which he purchases, or he does not raise it.

If he does not raise it, he gives no additional encouragement to production. The supposition, therefore, must be that he does raise prices. But exactly in proportion as he raises prices, he sinks the value of money. He therefore gives no additional encouragement to production" (1821, p. 123, as quoted in Corry, 1962, p. 40).

[3] On this point see Fetter (1942, pp. 369-71) who effectively refutes Viner's contention that Wheatley was concerned with the output effects of contraction. Also note that Ricardo's belief in money's neutrality extended only to the *level*, not the *composition*, of output. He (*Works*, I, 208-9) thought that, because the structure of excise taxes was fixed in nominal terms, money- and hence price-level changes could, via their effect on the real tax structure, alter profit rates and thus incentives to produce in different sectors of the economy. The result would be a change in the composition, though not the aggregate level, of output.

output and employment. Conversely, the lag of nominal wages behind prices would cause monetary contraction and the ensuing price deflation to raise real wages, lower real profits, and thereby discourage production and employment.

Henry Thornton was among the first to expound these points. He noted that declines in the stock of money would have no employment effect if wages fell as fast as prices. He then observed that wages in fact were downwardly inflexible in response to price falls, particularly temporary or unexpected ones. For that reason he thought monetary contraction would depress real activity. In his 1802 *Paper Credit* he wrote:

> It is true, that if we could suppose the diminution of bank paper to produce permanently a diminution in the value of all articles whatsoever and a diminution . . . in the rate of wages also, the encouragement to future manufactures would be the same, though there would be a loss on the stock in hand. The tendency, however, of a very great and sudden reduction of the accustomed number of bank notes, is to create an *unusual* and *temporary* distress, and a fall of price arising from that distress. But a fall arising from temporary distress, will be attended probably with no correspondent fall in the rate of wages; for the fall of price, and the distress, will be understood to be temporary, and the rate of wages, we know, is not so variable as the price of goods. There is reason, therefore, to fear that the unnatural and extraordinary low price arising from the sort of distress of which we now speak, would occasion much discouragement of the fabrication of manufactures (pp. 118-19).

Of Thornton's analysis two points are especially noteworthy. First, he attributes money-wage stickiness to the fact that wages are established on the basis of the expected long-run equilibrium price level which is much less volatile than temporary prices. In a long footnote attached to the preceding passage he explains that the equilibrium price level in an open economy operating under the gold standard is determined on purchasing-power-parity grounds by the given world gold price of goods. Second, he blames economic distress on *unexpected* contractions of the money stock. In so doing, he anticipates today's new classicals who argue that only unanticipated money matters for real variables.

To avoid deflation and its adverse effects, Thornton recommended preventing gold drains—particularly those arising from bank panics and/or real shocks to the balance of payments—from shrinking the money supply. The Bank of England should offset or sterilize such drains with compensating note issues, thus forestalling monetary contraction and its adverse consequences. He was even willing to risk temporary suspension of the gold standard rather than to let specie drains precipitate declines in the quantity of money. To him, suspension was preferable to contraction and the depression it would bring. He was equally opposed to inflation although he admitted that it could stimulate activity through the wage lag. Said he:

Henry Thornton (1760-1815)

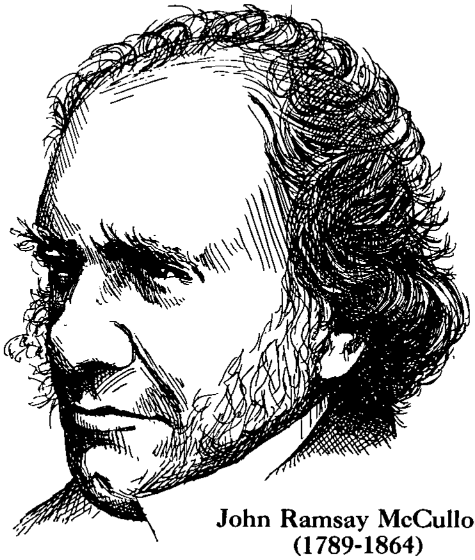
> . . . additional industry will be one effect of an extraordinary emission of paper, a rise in the cost [i.e., price] of articles will be another.
>
> Probably no small part of that industry which is excited by new paper is produced through the very means of the enhancement of the cost of commodities (p. 237).

Ricardo disagreed with Thornton. He did so on the grounds that wage flexibility rendered the lag too short for money to have more than a negligible impact on output. But other classicals concurred with Thornton. Among them was Robert Torrens who stressed the stimulus to profit and production emanating from sticky wages. When the Political Economy Club met in December 1830 to discuss Hume's theory of beneficial inflation, Torrens was in attendance to state his views. According to J. L. Mallet's account of the proceedings:

> Torrens . . . looks chiefly to profits as the great means of increasing general wealth, and as wages are fixed from time to time . . . and do not rise, perhaps for a long time after the value of money has fallen, the Capitalist pays in fact for long periods, lower real wages, and is a great gainer. All employers of Capital borrowed are likewise benefitted—paying less interest. There is a greater stimulus to production (Political Economy Club, 1921, p. 219, as quoted in Corry, 1962, p. 58).

Fixed Charges

Closely associated with sticky wages was another source of nonneutrality, namely the existence of contractually fixed costs, notably rents, taxes, and debt-service charges. Being fixed in nominal terms, these costs, the classicals explained, did not rise with prices, at least not in the short run. Consequently when prices rose due to monetary expansion the real burden of fixed costs fell. The corresponding rise in profits would spur output and employment. Conversely, monetary contraction and price deflation would, by raising the real burden of fixed nominal charges, discourage real activity.



John Ramsay McCulloch
(1789-1864)

Of the classical writers, J. R. McCulloch and Thomas Attwood stressed this particular source of nonneutrality. Thus O'Brien (1970), in his definitive study of McCulloch, writes that the latter saw the benefits of monetary inflation

> as being in reducing the weight of fixed burdens—rents and taxes—as they remained constant in money terms while the prices of final products increased, hence increasing profit margins. Increased profit stimulated production, employment, and wages. Precisely the opposite effect arose from reducing the quantity of money (pp. 160-61).

Thomas Attwood too held that rising prices spur activity by reducing the real burden of fixed costs or, what is the same thing, by increasing the gap between prices and these costs. "There is," he claimed, "no difficulty in employing and maintaining labourers, so long as the prices of the products . . . are *kept above the range of the fixed charges and monied expenses*" (1826, p. 42, italics in original). To him the extra profits arising from a widening of the gap between prices and fixed costs constituted the key to money's stimulus. "Prosperity," he wrote, has occurred whenever the government has

> filled the Country with what is called *Money*; and this *plenty of Money* has necessarily produced a general elevation of prices; and this general elevation of prices has necessarily produced a general increase of *profit* in all occupations; and this general increase of *profits* has, as a matter of course, given activity to every trade in the kingdom; and whilst the workmen, in one branch of trade, are *producing* one set of articles, they are inevitably *consuming* an equal amount of all other articles. This is the *prosperity of the Country*, and there is no other prosperity which ever has been enjoyed, or ever can be enjoyed (1826, pp. 11-12, italics in original).

Again,

> The . . . prosperity of the Country is indeed to be attributed to one cause only, and that cause is the general increase of the Circulating Medium (1826, p. 12).

By contrast, monetary contraction and deflation, he held, had the opposite effect. For when "*paper money* is withdrawn" and "the prices of commodities are suffered to fall . . . within the level of the *fixed charges and expences* . . . the industry of the country dies" (1826, p. 42, italics in original). It does so because "all the monied incumbrances," being fixed in nominal terms, "become encreased in real burthen, and operate in arresting all the means and the motives which conduce to the employment of labour, and to the production of national wealth" (1819, p. 42). Attwood concludes:

> When a [price] fall . . . takes place . . . first upon one article and then upon another, without any correspondent fall taking place upon debts and obligations, it has the effect of destroying all confidence in property, and all inducements to its production, or to the employment of laborers in any way (1817, pp. 78-79, as quoted in Viner, 1937, p. 186).

In short, owing to rigid cost elements, deflation leads to depression that brings suffering to the unemployed and distress to producers. It therefore follows, said Attwood, that

> it is the deficiency of money, and not its excess, which ought most to be guarded against, which produces want of employment, poverty, misery, and discontent in nations (1843, p. 18).

To prevent such disastrous monetary shortage he recommended that the Bank of England

> be obligated or otherwise be induced, to encrease the circulation of their notes as far as the national interests may require, that is to say, until all the labourers in the kingdom are again in full employment at ample wages (1819, p. 44).

To Attwood, full employment was the overriding policy goal and price increases the essential means of attaining it. Said he:

> so long as any number of industrious honest workmen in the Kingdom are out of employment, supposing such deficiency of employment not to be local but general, I should think it the duty, and certainly the interest, of Government, to continue the depreciation of the currency until full employment is obtained and general prosperity (1832, p. 467, quoted in Fetter, 1964, p. xxii).

Accordingly, "the great object of currency legislation should therefore be to secure and promote this gradual depreciation" (1817a, p. 101n, quoted in Checkland, 1948, p. 8). To this end he urged the government to

> Restore the depreciated state of the currency, and you restore the reward of industry, you restore confidence, you restore consumption, you restore every thing that constitutes the commercial prosperity of the nation (1816, p. 66).

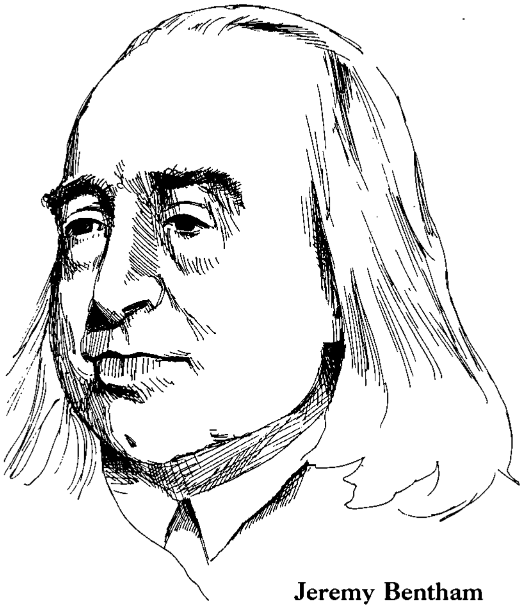
Attwood's inflationary policy views were too extreme even for other classical believers in the nonneutrality of money. John Stuart Mill (1833), for one, opposed Attwood's inflationism on the ground that it only works by tricking or deluding producers into thinking that nominal price changes are real and thus constitutes a deceitful and immoral way to stimulate activity. Mill did not, however, dispute Attwood's contention that inflation could raise profits by reducing the real burden of fixed costs. This item had become a standard element of the classicals' list of sources of nonneutrality.

Forced Saving

The classicals explained the fourth source of money's nonneutrality by means of their *forced-saving doctrine*.[4] The doctrine holds that monetary inflation stimulates capital formation and potential output by shifting real income from wage earners and fixed income recipients having high propensities to consume to capitalist entrepreneurs having high propensities to invest.

The doctrine originates with Jeremy Bentham who, assuming as he did continuous full employment, used it to argue that a monetary stimulus must operate through capital formation rather than through the activation of idle hands, as Hume had claimed. In his 1804 manuscript "Institute of Political Economy," the relevant parts of which were completed as early as 1800 or 1801, Bentham wrote:

> All hands being employed, and employed in the most advantageous manner, . . . the effect of every increase of money . . . is to impose an unprofitable *income tax* on the income of all fixed incomists.
>
> If . . . the additional money have come into hands by which it has been employed in the shape of capital, the suffering by the income tax is partly reduced and partly compensated. It is reduced by the mass of things vendible produced by means of it. . . . It is in a certain degree, though in a very inadequate degree, compensated for by the same means; viz. by the amount of the addition made to the quantity of sensible wealth—of wealth possessing a value in the way of use. Here . . . in the . . . case of forced frugality, national wealth is increased at the expense of national comfort and national justice (as quoted in Hayek 1932, p. 125).


Jeremy Bentham
(1748-1832)

Henry Thornton extended the doctrine when he argued that, owing to the lag of wages behind prices, forced saving could be extracted from wage-earners as well as from Bentham's fixed-income recipients. As he put it in his *Paper Credit*:

> Provided we assume an excessive issue of paper to lift up, as it may for a time, the cost [i.e., price] of goods though not the price of labour, some augmentation of stock will be the consequence; for the labourer . . . may be forced by his necessity to consume fewer articles, though he may exercise the same industry. But this saving, as well as any additional one which may arise from a similar defalcation of the revenue of the unproductive members of the society, will be attended with a proportionate hardship and injustice (p. 239).

Owing to these forced-saving effects, Thornton concludes that "paper possesses the faculty of enlarging the quantity of commodities by giving life to some new industry" (p. 239).

T. R. Malthus further elaborated the doctrine in his 1811 *Edinburgh Review* article on "Depreciation of Paper Currency." He held that forced saving was so potentially powerful in its effects on production that output could rise equiproportionally with the

[4] On the classicals' forced-saving doctrine see Hayek (1932) and Hudson (1965).

money stock leaving prices unchanged. Constituting the most complete description of the forced-saving mechanism in the classical literature, Malthus's statement warrants quotation in some detail. He starts by linking the money stock and its distribution to capital formation and real output.

> If such a distribution of the circulating medium were to take place, as to throw the command of the produce of the country chiefly into the hands of the productive classes . . . the proportion between capital and revenue would be greatly altered to the advantage of capital; and in a short time, the produce of the country would be greatly augmented (p. 96).

Thomas Robert Malthus
(1766-1834)

The key points, Malthus declares, are (1) that new money accrues to capitalists to raise the share of national income devoted to investment, and (2) that the corresponding required decrease in consumption is forced upon wage earners and fixed-income groups by the price rise caused by the monetary expansion. Thus

> A fresh issue of notes comes. . . . into the market, as so much additional capital, to purchase what is necessary for the conduct of the concern. But before the produce of the country has been increased, it is impossible for one person to have more of it, without diminishing the shares of some others. This diminution is effected by the rise of prices, occasioned by the competition of the new notes, which puts it out of the power of those who are only buyers, and not sellers, to purchase as much of the annual produce as before (p. 96).

From his analysis, Malthus concludes that

> On every fresh issue of notes, not only is the quantity of the circulating medium increased, but the distribution of the whole mass is altered. A larger proportion falls into the hands of those who consume and produce, and a smaller proportion into the hands of those who only consume. And as we have always considered capital as that portion of the national accumulations and annual produce, which is at the command of those who mean to employ it with a view to reproduction, we are bound to acknowledge, that an increased issue of notes tends to increase the national capital, and by an almost, though not strictly necessary consequence, to lower the rate of interest (pp. 96-97).

These effects, Malthus said, may explain why "a rise of prices is generally found conjoined with public prosperity; and a fall of prices with national decline" (p. 97).

Finally, Malthus notes that while forced saving necessarily operates through rising prices, the rise may be temporary. For

> it frequently happens, we conceive, that . . . the increased command of the produce transferred to the industrious classes by the increase of prices, gives such a stimulus to the productive powers of the country, that, in a short time, the balance between commodities and currency is restored, by the great multiplication of the former,—and prices return to their former level (pp. 97-98).

In terms of the equation of exchange MV = PQ, with velocity V constant, output Q rises to match the increase in money M leaving the equilibrium level of prices P unchanged.

Ricardo did not share Malthus's opinion of the productive power of forced saving. Though giving formal recognition to the doctrine, he denied its empirical importance. Thus he denied that redistribution from fixed-income receivers to capitalists could produce accumulation since both groups, he believed, possessed identical propensities to save and invest. In this case, he said, "there is a mere transfer of property, but no creation" of capital (*Works*, VI, 16). And while admitting the theoretical possibility that monetary expansion might extract forced saving from wage-earners via the lag of wages behind prices, he contended that wage flexibility in fact renders the lag too short and the resulting capital formation and output expansion too trivial to matter. Said he:

> There appears to me only one way in which any addition would be made to the Capital of a country in consequence of an addition of money; it would be this. Till the wages of labour had found their new level, with the altered value of money,—the situation of the labourer would be relatively worse; he would produce more relatively to that which he consumed, or rather would be obliged to consume less.

> The manufacturer would be enabled to employ more labourers as he would receive an additional price for his commodities; he might therefore add to his real capital till the rise in the wages of labour placed him in his proper sphere. In this interval some *trifling addition* would have been made to the Capital of the community (*Works*, VI, 16-17, emphasis added).

Likewise:

> There is but one way in which an increase of money . . . can augment riches, viz at the expence of the wages of labour; till the wages of labour have found their level with the increased prices . . . there will be so much additional revenue to the manufacturer . . . so that the real riches of the country will be somewhat augmented. A productive labourer will produce something more than before relatively to his consumption, *but this can be only of momentary duration* (*Works*, III, 318-19, emphasis added).

In sum, Ricardo, unlike the other classicals, was extremely skeptical of the forced-saving idea.

Although the above economists disputed the size of forced saving's effects, none disputed the distributive injustice involved. All saw forced saving as an immoral and deceitful means of stimulating accumulation and on that ground condemned its use.

Not so J. R. McCulloch, however. He praised forced saving and its inflationary effects and rejected any considerations of injustice. He readily acknowledged that inflation shifts real purchasing power from fixed-income consumers to capitalist investors. But unlike the others, he lauded such redistribution on the grounds that the gainers exceeded the losers. Besides entrepreneurs, the gainers included the whole community which benefited from increased output, employment, and capital formation. The losers were confined to a small group of rentiers and annuitants but excluded wage-earners since wages, he felt, tended to rise with prices. The losers' suffering he thought a small price to pay for the general benefits of inflation.[5] Thus, at the December 3, 1830 meeting of the Political Economy Club, he callously dismissed Thomas Tooke's solicitude for fixed-income recipients. According to J. L. Mallet's Diaries:

> McCulloch in his sarcastic and cynical manner derided Mr. Tooke's concern for old gentlemen and ladies, dowagers, spinsters and land holders. He cared not what became of them, and whether they were driven from the parlour to the garret, provided the producers—the productive and industrious classes—were benefited, which he had no doubt they were by a gradual depreciation in the value of money (Political Economy Club, 1921, p. 219, as quoted in O'Brien, 1970, p. 166).

Although he extolled inflation, McCulloch's main concern was with the evils of deflation. In this connection he argued that any ill effects of paper money expansion came not from inflation per se but from the eventual need to contract to protect the nation's gold reserve. He feared that the damage wreaked by the resulting deflation would far exceed the gains from the preceding inflation. As proof, he noted that the prosperity associated with inflation during the Napoleonic Wars was more than offset by the distress that accompanied the deflation in the immediate postwar period. To him, avoiding monetary contraction was far more important than promoting monetary expansion. His emphasis on the damage of deflation was typical of classical believers in the short-run nonneutrality of money.

Confusion of Monetary for Real Shocks

The classicals traced a fifth source of nonneutrality to a confusion between general and relative prices. They explained that money has real effects because changes in its quantity cause general price movements which producers mistake for real relative price changes requiring output adjustments. Fooled by unexpected monetary growth and the resulting economy-wide rise in prices, economic agents treat the price increases as signifying demand shifts special to themselves and so expand production.

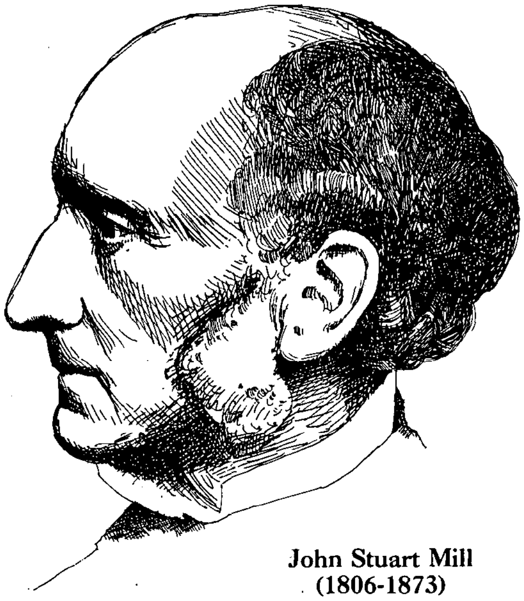
Credit for identifying this particular nonneutrality goes to John Stuart Mill. In his 1833 article "The Currency Juggle," he explained how unanticipated money growth had

> produced a rise of prices, which *not* being supposed to be connected with a depreciation of the currency, each merchant or manufacturer considered to arise from an increase of the effectual demand for his particular article, and fancied there was a ready and permanent market for almost any quantity of that article which he could produce (p. 191).

In other words, each producer had misinterpreted the rise in general prices as a relative-price signal to expand his operations. Here is how monetary expansion and the resulting general inflation may, in Mill's words, "create a *false opinion* of an increase of demand, which false opinion leads, as the reality would do, to an increase of production . . ." (p. 191).

Mill recognized that the confusion between general and relative prices applies equally to workers who,

[5] Torrens in his 1812 *Essay on Money and Paper Currency* took much the same position. He wrote that fixed-income receivers constitute "so small a proportion to the whole community, that any inconvenience they may suffer, from a fall in the value of money, sinks into insignificance, nay entirely vanishes, when compared with the universal opulence, the general diffusion of happiness arising from augmented trade, and the rise in the wages of labour, which the increased quantity of money is instrumental in producing" (pp. 40-41, as quoted in Robbins, 1958, p. 76).



John Stuart Mill
(1806-1873)

failing to see that price rises are so extensive as to reduce real wages, supply extra effort under the misapprehension that nominal wage increases constitute real ones. He explains:

> the inducement which . . . excited this unusual ardour in all persons engaged in production, must have been the expectation of getting more commodities generally, more real wealth, in exchange for the produce of their labour, and not merely more pieces of paper (1848, p. 550).

Mill was no believer in long-run nonneutrality. He insisted (1) that inflation's stimulus is temporary at best, (2) that it lasts only "as long as the existence of depreciation is not suspected" or anticipated (1844, p. 275), (3) that it ends "when the delusion vanishes and the truth is disclosed" (1844, p. 275), and (4) that it is "followed . . . by a fatal revulsion as soon as the delusion ceases" (1833, p. 191). In other words, once agents correctly perceive wage and price increases as nominal rather than real, economic activity reverts to its steady-state level, but only after undergoing a temporary recession to correct for the excesses of the inflationary boom. Here is Mill's conclusion that, when people mistake general for relative price increases, nonneutrality arises both at the time of the misperception and also when it is corrected. Mill's insistence that only unperceived or unanticipated inflation has real effects marks him as a forerunner of the modern new classical school.

Other Sources of Nonneutrality

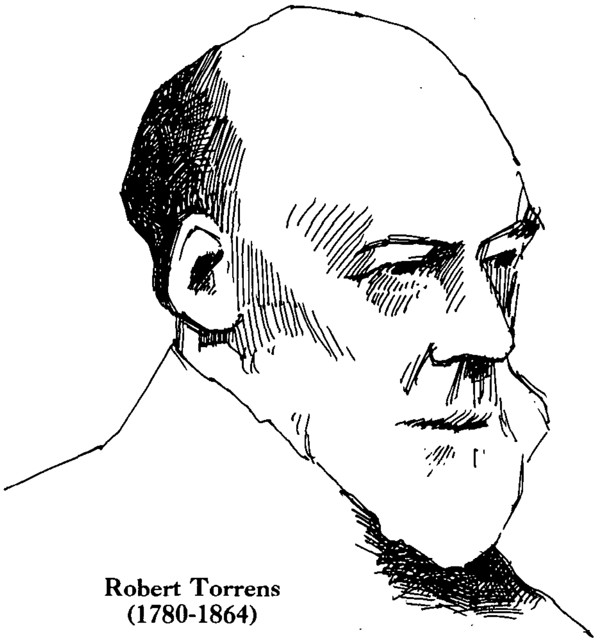
The preceding by no means exhausts the list of nonneutralities considered by the classicals. Also analyzed were at least four more.

The first relied on Adam Smith's doctrine that the division of labor is limited by the extent of the market. Attwood, Malthus, McCulloch, and Torrens employed this idea. They argued that monetary expansion stimulates aggregate spending which enhances the scope of the market for goods and services. In Attwood's words:

> the issue of money *will* create markets, and . . . it is upon the abundance or scarcity of money that the extent of all markets principally depends (1817b, p. 5, as quoted in Fetter, 1965, p. 75).

Similarly Torrens claimed that extra money improves business confidence and that "an enlargement of confidence always produces that enlargement of the market which it anticipates" (1816, as quoted in Robbins, 1958, p. 82). Extension of the market then prompts increased specialization and division of labor, thus boosting labor's productivity. Through this channel monetary expansion, in Torrens's words, "facilitates exchanges, and, by occasioning more accurate division of employment, augments the productiveness of industry" (1812, p. 95, as quoted in Robbins, 1958, p. 77). In so doing, money growth induces a higher level of output from a given labor force.[6]

[6] Traces of the division-of-labor argument survive today in the popular notion that scale economies enable firms to respond to demand-expansion policy by producing higher levels of output at lower unit costs.



Robert Torrens
(1780-1864)

Nor is this all. For Torrens in particular recognized that the labor force itself might expand under the impact of inflationary money growth. He thought that rising prices, by eroding the real value of fixed nominal incomes, could force annuitants, rentiers, and the like to go to work in an effort to maintain their real incomes. Such people, he said,

> finding their places in society perpetually sinking, will be prompted to some species of exertion, in order to avert the evil; and thus the number of idle individuals, who add nothing to the general stock of society, will be diminished, and industry will receive a two-fold stimulus,

namely one arising from increased division of labor and the other from augmentation of the labor force (1812, pp. 40-41, as quoted in Robbins, 1958, p. 76).

Torrens also acknowledged that money growth could stimulate industry if nominal interest rates lagged behind inflation so that real rates fell. He said that when this happens "all employers of Capital borrowed are likewise benefitted—paying less [real] interest. There is a greater stimulus to production" (Political Economy Club, 1921, p. 219, as quoted in Corry, 1962, p. 58).

Division of labor, expansion of the labor force, lag in nominal interest rate—these constituted three of the four additional sources of nonneutrality identified by the classicals. Henry Thornton located the fourth in sellers' efforts to maintain constant real inventory-to-sales ratios. These efforts, which ensured that any money-induced rise in the real volume of sales would be matched by a corresponding rise in production for inventory, were described by him as follows:

> It may be said . . . that an encreased issue of paper tends to produce a more brisk demand for the existing goods, and a somewhat more prompt consumption of them; that the more prompt consumption supposes a diminution of the ordinary stock, and the application of that part of it, which is consumed, to the purpose of giving life to fresh industry; that the fresh industry thus excited will be the means of gradually creating additional stock, which will serve to replace the stock by which the industry had been supported; and that the new circulating medium will, in this manner, create for itself much new employment (1802, p. 237).

All-in-all the classicals left a fairly extensive list of factors explaining money's short-run output effects.

The Classicals' Legacy

The classicals bequeathed their theory of nonneutrality to later generations of economists who used it to account for money's temporary impact on real variables. Thus quantity theorists from Irving Fisher to Milton Friedman introduced Hume's price lag into the equation of exchange MV = PQ to show that, with velocity V constant, a change in the money stock M produces a temporary change in output Q before fully changing prices P.[7] Keynesians employed the same notion to argue that, with unemployed resources, prices fail to rise in proportion with a rising nominal money stock. The resulting rise in the real money stock, Keynesians claimed, lowers the rate of interest and thereby boosts investment spending and thus the level of national income.[8]

Other classical sources of nonneutrality were quickly absorbed into mainstream monetary thought. Alfred and Mary Marshall (1879, pp. 155-56), A. C. Pigou (1913, pp. 75-84), Ralph Hawtrey (1913), and Keynesians in the 1940s, '50s, and '60s used the notion of sticky money wages to explain how fluctuations in prices caused or accommodated by fluctuations in money produce corresponding fluctuations in real wages and thus output and employment. Irving Fisher (1913, Ch. 4) employed the idea of sticky nominal interest rates to explain how money-induced price changes affect investment and real activity by changing real rates. This idea formed the basis of his (1923) theory of the business cycle as "a dance of the dollar." Likewise his (1933) debt-deflation theory of the Great Depression embodied the classical idea that falling prices emanating from monetary contraction depress real activity by raising the real burden of debt-service charges.

Additional classical ideas were put to work. Austrian economists Ludwig von Mises (1912) and Frederich von Hayek (1933) used the classical doctrine of forced saving to explain the upswing phase of their monetary overinvestment theory of the cycle. And most recently, Robert Lucas (1972) has developed John Stuart Mill's idea that money has real effects when general price changes are mistaken for relative price ones. Also prominent in Lucas's and other new classicals' analysis is the Thornton-Mill argument that real effects stem from ***unanticipated*** money. Classical contributions are thus seen to underlie much twentieth-century work on money's nonneutrality.

These contributions notwithstanding, the myth persists that the classicals adhered to the neutrality

[7] On the nonneutrality of money in the writings of Irving Fisher, the Chicago school, and the Cambridge cash-balance school, see Patinkin (1972).

[8] See Patinkin (1987, p. 640).

proposition in the short run as well as the long. Keynes created this myth in his *General Theory* when he sought to differentiate his approach from those of his classical and neoclassical predecessors. Today economists and textbook writers perpetuate the myth by disseminating a caricature "classical" macromodel in which money is always neutral. Further contributing to the myth is the tendency of writers such as Arjo Klamer (1984, p. 12) to interpret the new classical macroeconomics and its policy-ineffectiveness idea as a return to an original classical tradition of neutrality-always. All are wrong. The classical tradition never held that money was always neutral. On the contrary, except for Ricardo and one or two others, the classicals believed that money had powerful temporary real effects and perhaps some residual permanent effects as well. In the view of the classicals, nonneutrality typified the short run and neutrality at best held approximately in the long run only.

References

Artis, Michael J. 1984. *Macroeconomics.* Oxford: Clarendon Press.

Attwood, Thomas. 1964. *Selected Economic Writings of Thomas Attwood,* ed. F. W. Fetter, London: LSE Reprints of Scarce Works on Political Economy, 1964.

————. 1816. *The Remedy; or, Thoughts on the Present Distresses.* London. As reprinted in his *Selected Economic Writings.*

————. 1817a. *Prosperity Restored; or, Reflections on the Cause of the Present Distresses and on the Only Means of Relieving Them.* London.

————. 1817b. *A Letter to the Right Honorable Nicholas Vansittart.* Birmingham.

————. 1819. *A Letter to the Earl of Liverpool.* Birmingham. As reprinted in his *Selected Economic Writings.*

————. 1826. *The Late Prosperity, and the Present Adversity of the Country, Explained.* London. As reprinted in his *Selected Economic Writings.*

————. 1832. *Report from the Committee on Secrecy in the Bank of England Charter,* Parliamentary Papers (Commons) 1831-32, vi, Q 5758.

————. 1843. *Thomas Attwood's Letter to Sir Robert Peel on the Currency.* As reprinted in his *Selected Economic Writings.*

Bentham, Jeremy. 1801-4. "The Institute of Political Economy." In Vol. III of *Jeremy Bentham's Economic Writings,* ed. W. Stark, London: George Allen & Unwin, 1954.

Checkland, S. G. 1948. "The Birmingham Economists," 1815-1850. *The Economic History Review,* 1-19.

Corry, B. A. 1962. *Money, Saving and Investment in English Economics: 1800-1850.* London: Macmillan.

Fetter, Frank W. 1942. "The Life and Writings of John Wheatley." *Journal of Political Economy* 50, June, 357-76.

————. 1964. "Introduction." *Selected Economic Writings of Thomas Attwood.* London.

————. 1965. *Development of British Monetary Orthodoxy 1797-1875.* Cambridge, Harvard University Press.

Fisher, Irving. 1913. *The Purchasing Power of Money.* Revised ed. New York: Macmillan. Reprinted, New York: Kelley, 1963.

————. 1923. "The Business Cycle Largely a 'Dance of the Dollar.' " *Journal of the American Statistical Association* 18, December, 1024-28.

————. 1933. "The Debt-Deflation Theory of Great Depressions." *Econometrica* 1. October, 337-57.

Friedman, Milton. 1970. *The Counter-Revolution in Monetary Theory.* London: Institute of Economic Affairs.

Glasner, David. 1989. *Free Banking and Monetary Reform.* New York: Cambridge University Press.

Hawtrey, Ralph. 1913. *Good and Bad Trade.* London: Constable.

Hayek, Frederich A. von. 1932. "A Note on the Development of the Doctrine of 'Forced Saving.' " *Quarterly Journal of Economics* 47, November, 123-33.

————. 1933. *Monetary Theory and the Trade Cycle.* London: Jonathan Cape.

Hoover, Kevin D. 1988. *The New Classical Macroeconomics: A Skeptical Enquiry.* New York: B. Blackwell.

Hudson, M. A. 1965. "Ricardo on Forced Saving." *Economic Record* 41, June, 240-47.

Hume, David. 1752. "Of Money" and "Of Interest." In D. Hume, *Writings on Economics,* ed. E. Rotwein, Madison: University of Wisconsin Press, 1970.

Klamer, Arjo. 1984. *Conversations with Economists: New Classical Economists and Opponents Speak Out on the Current Controversy in Macroeconomics.* Totowa, NJ: Rowman & Allanheld.

Lucas, Robert E. 1972. "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4, April, 103-24.

Malthus, T. R. 1811. "Depreciation of Paper Currency." *Edinburgh Review* 17, February, 340-72. In *Occasional Papers of T. R. Malthus,* ed. B. Semmel, New York: Burt Franklin, 1963, 71-104.

Marshall, Alfred, and Mary P. Marshall. 1879. *Economics of Industry.* London: Macmillan.

Mill, James. 1821. *Elements of Political Economy*. London: Baldwin, Craddock, and Joy.

Mill, John Stuart. 1833. "The Currency Juggle." *Tait's Edinburgh Magazine* 2, January, 461-67. As reprinted in Vol. IV of *The Collected Works Of John Stuart Mill*, ed. J. M. Robson, Toronto: University of Toronto Press, 1967, 181-92.

————. 1844. "Of the Influence of Consumption on Production." In *Essays on Some Unsettled Questions of Political Economy*. London. As reprinted in Vol. IV of *The Collected Works of John Stuart Mill*, ed. J. M. Robson, Toronto: University of Toronto Press, 1967, 262-79.

————. 1848. *Principles of Political Economy with Some of Their Applications to Social Philosophy*. 7th ed. 1871. As reprinted in the Ashley edition, ed. W. T. Ashley, London: Longmans, Green, and Co., 1909.

Mises, Ludwig von. 1912. *The Theory of Money and Credit*. 2d. ed. 1934. London: Jonathan Cape.

Niehans, Jürg. 1987. "Classical Monetary Theory, New and Old." *Journal of Money, Credit, and Banking* 19, November, 409-24.

O'Brien, D. P. 1970. *J. R. McCulloch: A Study in Classical Economics*. New York: Barnes and Noble.

————. 1975. *The Classical Economists*. Oxford: Clarendon Press.

Papademos, Lucas, and Franco Modigliani. 1990. "The Supply of Money and the Control of Nominal Income." Chapter 10 of Benjamin Friedman and Frank Hahn, eds., *Handbook of Monetary Economics* 1, New York: North-Holland, 399-494b.

Patinkin, Don. 1972. "On the Short-Run Non-Neutrality of Money in the Quantity Theory." *Banca Nazionale del Lavoro Quarterly Review* 100, March, 3-22.

————. 1987. "Neutrality of Money." In *The New Palgrave: A Dictionary of Economics*, ed. J. Eatwell, M. Milgate, P. Newman.

Pigou, Arthur C. 1913. *Unemployment*. London: Williams and Norgate.

Political Economy Club. 1921. *Proceedings*, Vol. 6, Centenary Volume. London.

Ricardo, David. 1951-73. *The Works and Correspondence of David Ricardo*, ed. P. Sraffa, 11 Vols. Cambridge.

Robbins, Lionel C. 1958. *Robert Torrens and the Evolution of Classical Economics*. London. Macmillan.

Thornton, Henry. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britian*, ed. F. A. v. Hayek. London: Allen & Unwin, 1939.

Torrens, Robert. 1812. *Essay on Money and Paper Currency*.

————. 1816. Letter to the *Sun* Newspaper. April 23.

Viner, Jacob. 1937. *Studies in the Theory of International Trade*. New York: Harper.

Wheatley, John. 1803. *Remarks on Currency and Commerce*. London.

Productivity in Banking and Effects from Deregulation*

David B. Humphrey

I. INTRODUCTION

There has been a marked decrease in the rate of productivity growth in the United States and other countries since the early 1970s. The likely reasons for this slowdown have been surveyed recently in Cullison (1989). The slowdown shows up in measures of single factor (labor) productivity as well as in the more comprehensive multifactor measure, which includes the productive effects of labor and capital together. For example, productivity in the U.S. nonfarm business sector only rose at a 0.22 percent annual average rate over 1973-87. But for the 25 years prior to 1973, productivity growth was over seven times larger (at 1.68 percent a year). The slowdown was even more striking for some U.S. service sectors. In particular, the Finance, Insurance, and Real Estate (FIRE) service sector experienced an average labor productivity growth rate that was negative, at −0.41 percent a year over 1973-87. In the 25 years before 1973, however, this growth averaged 1.41 a year (Baily and Gordon, 1988, pp. 355, 395).

Banking makes up 20 percent of the FIRE service sector (net of owner-occupied housing) and thus contributes importantly to this sector's behavior. The purpose of this paper is to provide estimates of total factor productivity for the banking service sector over the past decade (1977-87) and to investigate the cause of the low productivity growth found. Productivity results are reported from two growth accounting models: one based on a production function and another based on a cost function. Both approaches indicate a similarly low rate of productivity advance for the banking industry, ranging between −0.07 (production approach) to 0.6 percent (cost approach) a year.

It is argued that low productivity growth in banking is largely due to the effects of bank deregulation initiated in the early 1980s. Deregulation permitted the establishment of new interest-bearing consumer checking accounts and eliminated ceilings on time and savings deposit interest rates. Deregulation during the 1980s, preceded by the intensive use of cash management techniques by corporations in the 1970s, effectively removed banks' virtual monopoly control over zero-interest checking accounts and low-interest small consumer time and savings deposits. Core deposit interest costs rose but were not offset by either reduced costs elsewhere or with an expansion in measured bank output. Apparently, market share considerations limited the desire by banks to reduce operating costs enough to fully offset the rise in interest expenses.

While banks may have experienced very low (to negative) productivity growth, users of banking services have benefited. But the benefits, which are similar to an increase in the "quality" of banking output, are not captured in any measure of banking output. Thus, although measured bank productivity growth is low or negative, it would be inappropriate to conclude that society as a whole has not benefited. Rather, there has been a redistribution of productivity benefits in which users of banking services have gained at the expense of banks.

II. PRODUCTIVITY IS "OUTPUT PER UNIT OF INPUT," BUT WHAT IS BANK OUTPUT AND WHAT ARE THE INPUTS?

What Do Banks Produce?

In many industries, physical measures of output and inputs are readily available and, importantly, a consensus also exists on how best to measure them. In the electric power industry, for example, the obvious measure of output is kilowatt-hours of electricity produced. Inputs used to produce electric power include the number of workers, the real value

* The opinions expressed are those of the author alone. Comments by William Cullison, Tony Kuprianov, and David Mengle are appreciated. Alex Wolman contributed outstanding research assistance.

of electric generators and transmission facilities, and the tons of fuel inputs used. In contrast, in the banking sector physical measures of output are not readily available (although they exist for some banks); indeed no strong consensus exists regarding what it is that banks produce. As a result, measures of banking productivity can use different definitions of outputs and inputs.

Banks produce a variety of payment, safekeeping, intermediation, and accounting services for deposit and loan customers (Benston and Smith, 1976; Mamalakis, 1987). Some have argued, however, that banks primarily produce loans. With this (asset) approach, the production of deposit services is viewed as merely payment in kind for the use of funds from which to make loans (Sealey and Lindley, 1977). In effect, this is a "reduced form" model of the banking firm: the production of deposit services is treated as an intermediate output to depositors who provide loanable funds, so deposit services are netted out.

But there is no reason to focus on only a single banking output such as loans, especially because the production of deposit services accounts for half of all physical capital and labor input expenditures. Because deposit services are such a large component of bank value added, explicit modeling of their productive structure, along with that of loans, will yield a more accurate description of this structure for the bank as a whole. This objective can be achieved using a structural model of a multiproduct banking firm. In such a model, the production of deposit services would not be netted out; instead, it would be one of a set of bank outputs.

For purposes of analysis, banks are considered to produce payment and safekeeping outputs (associated with demand deposits and savings and small denomination time deposits) as well as intermediation and loan outputs (associated with real estate loans, consumer installment and credit card loans, and commercial, industrial, and agricultural loans). Over the last decade, these five deposit and loan output categories accounted for 75 to 80 percent of value added in banking (Berger and Humphrey, forthcoming, see table). Such a categorization of bank output, with one exception (time deposits), is consistent with that identified in the user cost approach to determine bank inputs from outputs (Hancock, 1986; Fixler and Zieschang, forthcoming).

Measures of Bank Output

Based on data availability, there are at least three different measures of banking output that could be used in productivity analyses: (1) the number of transactions processed in deposit and loan accounts (a flow measure); (2) the real or constant dollar value of funds in the deposit and loan accounts (a stock measure); or (3) the numbers of deposit and loan accounts serviced by banks (a stock measure).[1] Because output is typically a flow, not a stock, the preferred measure is seemingly an output flow. Stock measures would only be used if a flow measure were unavailable or because the stock measure might be proportional (on average) to a flow measure.

A time-series transactions flow measure of aggregate banking output is compiled by the Bureau of Labor Statistics (BLS, 1989). However, this measure exists only for the aggregate of all banks and has a limited number of observations. Thus for most purposes, researchers have been forced to rely on stock measures of bank output and to assume that there is a proportionality between stocks and flows, so use of stocks succeeds in approximating flows. Because one possible stock measure—number of deposit and loan accounts—is essentially unavailable for time-series analysis,[2] researchers have relied on the stock

Summary of Bank Total Factor Productivity Estimates

(annual average growth rates; 1977-87)

	QT	QD
Growth Accounting Method:		
Production Function	-0.00%	-0.07%
Cost Function	0.60	0.50
Econometric Estimation Method:[1]		
Cost Function:		
Hunter & Timme (1991)	—	1.05
Humphrey (1991)	—	-1.01

[1] Both of these studies used multiproduct indicators of bank output rather than the single aggregate index QD. Transactions flow data (QT) are not available to be used in pooled times-series, cross-section econometric analyses.

[1] A fourth measure, concerning bank debits and deposit turnover (published monthly in the *Federal Reserve Bulletin*), should not be used. These data are in value terms and include both check and wire transfer debits. As a result, the virtually exponential growth in the value of wire transfers will grossly dominate this series, even though wire transfer expenses are a minute portion of total bank costs. While it is possible to remove the value of wire transfer debits, the end result would be a measure of the value of check and ACH debits, which is inferior to the quantity measure of aggregate check and ACH transactions captured in the transaction flow measure discussed immediately below.

[2] See the Appendix for more detail on data availability.

of the real value of deposits and loans. These data are available over time and for each bank in the United States. As a result, cross-section information can be pooled over time, allowing the estimation of more sophisticated econometric models than is possible with any of the other measures of bank output. It is assumed, but has never been tested, that the transaction flow of bank output over time is proportional to the stock of real deposit and loan balances (Box 1).[3] That these two alternative measures of bank output have had a somewhat similar variation over the last decade is documented below. While this does not strongly support the assumption of strict proportionality between bank output flow and stock, it does suggest that somewhat similar estimates of productivity may be obtained using either output measure for this period. This point is demonstrated below.

Box 1

When Will Stock and Flow Measures of Bank Output Be Proportional to Each Other?

Stock and flow measures of banking output will be proportional to one another when only the two following influences determine the growth in nominal deposit and loan balances over time. First, nominal deposit and loan balances grow because of population growth. An expanding population leads to a larger demand for bank transaction services as more deposit accounts are opened, more checks are written, and more savings deposits and withdrawals occur. Thus, over time, increased transaction flows will be associated with larger stocks of deposit balances. Population growth and economic expansion also leads to loan growth. The nominal value of the stock of bank loans will rise as new loan transactions occur and expand at a greater rate than outstanding loans are retired. The second influence is inflation, which raises the average size of loans made and the average idle deposit balances held by users of bank services. If only these two influences determine the variation in nominal deposit and loan balances, then deflation by some appropriate price index will give the real value of deposit and loan balances and also reflect the underlying flow of bank transactions.

Inputs Needed to Produce Output

There is less controversy on measuring bank inputs. Labor (number of workers or total hours worked) and the real or constant dollar value of physical capital (usually the book value of premises, furniture, and equipment deflated by some price index) clearly represent inputs needed to produce bank output.[4] However, there is less agreement about also treating the real or constant dollar value of loanable funds—core deposits plus purchased funds—as an input.

If labor and capital were the only inputs, then measured productivity would refer to bank operating costs. Since operating costs are less than one-third of total banking costs, however, an operating cost productivity measure by itself would not indicate the degree to which productivity improvements may affect user costs or bank profits. More importantly, since capital and labor operating expenses which support a branch network are substitutes for the interest costs of purchased funds (federal funds, CDs, Eurodollars, etc.), operating expenses are not a stable proportion of total costs either over time or (especially) across different-sized banks.[5] This instability can bias productivity estimates derived solely from operating expenses, just as it has been shown to bias the determination of bank scale economies (Humphrey, 1990). Hence the appropriate cost concept from which to estimate bank productivity is total costs, which includes operating plus interest expenses. From this it follows that the five appropriate inputs are labor, capital, demand deposits, small time and savings deposits, and purchased funds. Thus a total factor measure of productivity is preferred.

Unlike other industries, total costs for an aggregate bank cannot be determined by simply summing all costs at all banks. Some costs, such as the cost of funds purchased from other banks in the interbank

[3] The same assumption is made in cross-section studies in banking where scale economies are the focus of modeling and estimation.

[4] Researchers familiar with the many problems associated with measuring real capital stock will find the measurement method employed in this paper to be overly simple and potentially misleading. Fortunately, these capital measurement problems will have only a relatively small effect on the banking productivity results because the share of capital expenditures in total cost is itself small, around 15 percent.

[5] Purchased funds permit a bank to grow faster and attain a larger size than if it relied solely on a base of branch-generated deposits.

funds market (e.g., federal funds), are costs only to individual banks but need to be excluded when aggregate data are used. This exclusion is necessary because if there were only one aggregate bank, which is the implicit assumption in using aggregate data in the type of models specified, interbank costs would not exist and total costs need to be reduced by this amount. The cost of funds purchased outside of the U.S. banking system, such as virtually all large CDs, Eurodollars, and other liabilities for borrowed money, however, would remain.

To sum up, both input (cost) and output (service flow or stock) characteristics of core deposits are specified (following Wykoff, 1991), rather than only one or the other as is usually done in the literature. In contrast, purchased funds have only input characteristics. Overall, five categories of bank output and five areas of input costs are specified.

III. GROWTH ACCOUNTING ESTIMATES OF BANKING PRODUCTIVITY

There are essentially two ways to measure bank productivity. The growth accounting approach (Box 2) uses raw data on input and output growth rates plus information on input cost shares while an econometric approach specifies a cost or production function relating outputs to inputs and estimates this relationship statistically. While the focus in this paper is on the growth accounting approach, results of existing econometric studies of bank technical change and productivity are also noted.

The data necessary to determine banking productivity from growth accounting models based first on a production function and second on a cost function (both shown in Box 2) are different with the exception of the measure of bank output. In what follows, the time-series variation of two bank output measures are compared, after which productivity results based on these output measures in both production and cost-growth accounting models are then contrasted.

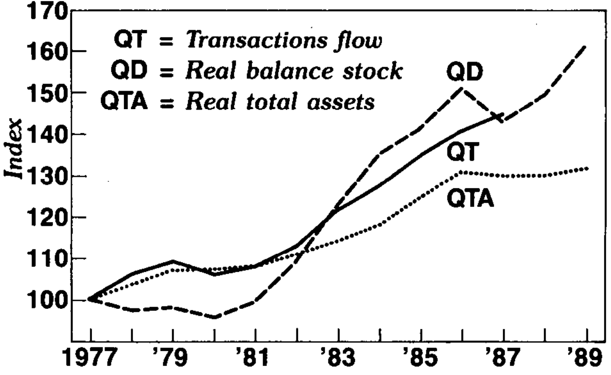
Transactions Flow and Real Balance Stock Measures of Bank Output

The transaction measure of bank output used here is the BLS index of deposit and loan transactions (QT). In contrast, the stock measure is an index of the real value of deposit and loan account balances (QD).[6] Both are shown in Figure 1. For comparison purposes, the real value of total bank assets (QTA) is also shown.[7] Over 1977-87, the annual average rate of growth of QT was 3.8 percent while that for QD was almost identical at 3.7 percent. But the average figures can be misleading since QD was very flat in the early 1980s but grew more rapidly than QT at the middle of the decade. Thus the assumed proportionality between bank transactions flows (QT) and the stock of real balances (QD) is only approximate over this period even though the R^2 between QT and QD is relatively high (.82). In comparison, QTA grew by only 2.7 percent on an annual average basis and, if used as a measure of banking output here (as some have argued), would understate the expansion of bank output compared with the other two measures.[8] Such understatement holds even though the R^2 between QT and QTA is higher (.97) than that between QT and QD.

Figure 1

A Comparison of Flow and Stock Measures of Banking Output

(1977-87 or 89; 1977=100)



A Production-Based Measure of Banking Productivity

The Bureau of Labor Statistics computes annually an aggregate measure of labor productivity in

[6] The construction of both of these indexes are described in the Appendix. The BLS data are available only through 1987 (BLS, 1989).

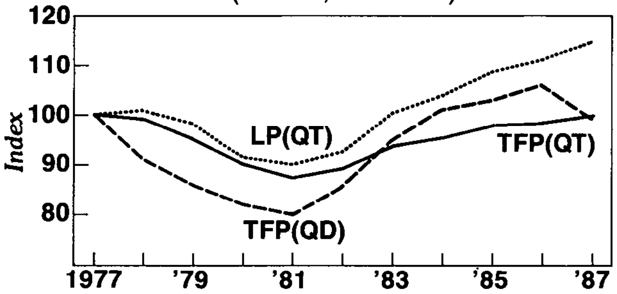
[7] Real total assets were obtained by deflating the nominal value of total banking assets by the GNP deflator.

[8] Since interbank sales of funds (e.g., federal funds sold) have grown over time and show up in total assets, the aggregate value of these assets will be overstated by this amount compared to a situation where there is only one aggregate bank and interbank sales no longer appear on the balance sheet. Thus the understatement possible when using total assets as an indicator of aggregate bank output is even greater than that shown in the figure since these total asset values have not been corrected for this double counting.

Figure 2

Production Approach: Single-Factor (Labor) and Total Factor Productivity

(1977-87; 1977=100)



TFP(QT) = *Total factor productivity with transactions flow output*

TFP(QD) = *Total factor productivity with real balance stock output*

LP(QT) = *Labor productivity with transactions flow output*

banking using transactions (QT) as its measure of output. This series, LP(QT), is shown in Figure 2. Cyclical behavior of labor productivity is due to cycles in bank output transactions flows, specifically cycles in new loans being made as deposit transaction growth was always positive.[9]

Over the 1977-87 period, the average annual increase in numbers of workers was 2.4 percent[10] while banking output (QT) rose by an average 3.8 percent. Because output grew faster than the labor input, labor productivity is positive (at 1.4 percent a year). But labor productivity is not representative of overall banking productivity if other inputs grew more rapidly or slowly than labor.[11]

Our (rough) estimate of the growth of the real value of bank physical capital is 1.8 percent annually with the real value of demand deposits falling by 3.5 percent, time and savings deposits growing by 5.9 percent, and purchased funds growing by 3.1 percent.[12] The net result is that the cumulative level of total factor productivity (TFP), using the QT transactions flow output measure, is below that for labor productivity. A similar result occurs when TFP is derived using the QD real balance stock output measure. Overall, neither measure of total factor productivity in a production-based growth accounting model shows any growth[13] while the BLS labor productivity measure grows by 1.4 percent a year.[14]

A Cost-Based Measure of Banking Productivity

In a cost-based growth accounting approach (see Box 2), input prices are used in place of input quantities and costs are attached to producing bank output. The productivity results using both output measures in a cost model are shown in Figure 3. While the time pattern of the productivity indexes differ over 1977-87, they start and end at almost the same points so their annual average growth rates are again quite similar, only this time they are slightly positive—a 0.6 percent growth rate for QT and 0.5 percent for QD.[15]

The differences in productivity estimates between the production and cost approaches can be seen in Figure 4. Total factor productivity estimates

[9] This result is seen in unpublished data on the six separate components of QT (described in the Appendix) from the BLS.

[10] Real labor input is from the BLS series on number of workers in banking. The number of full-time equivalent workers from the *Call Report* grew by only 1.6 percent a year over the same period.

[11] The bank labor productivity series derived in Baily and Gordon (1988), p. 395, cannot be used for comparison here. This is because their measure of bank output growth, derived from National Income and Product Account data, is itself based on the growth of the labor input. Thus labor productivity growth will be zero by definition as the growth in bank output equals that of the labor input.

[12] The real value of these three funds categories is the nominal value divided by the GNP deflator. The real value of bank capital is described in the Appendix.

[13] More specifically, TFP using QT (QD) in the production-based growth accounting model has a growth rate of −0.0 (−0.07) percent. The difference in TFP using QD versus QT is directly related to QD being flat in the late 1970s but experiencing more rapid growth than QT in the mid-1980s (see Figure 1).

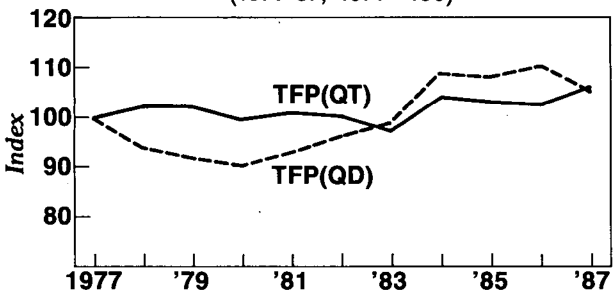
[14] Two alternative deflators for the replacement price of bank physical capital were used for illustration. These were the GNP deflator and the ratio of current capital expenditures (historical depreciation) to the book value of physical capital. For the QT output measure, average annual TFP was −0.28 percent and −0.58 percent, respectively (rather than −0.0 percent as reported above). For the QD output measure, these rates were −0.35 percent and −0.64 percent (rather than −0.07 percent as reported). All of these results use the BLS series on the number of banking workers rather than the (slower growing) number of full-time equivalent workers from the *Call Report*. Use of the *Call Report* labor data would change the QT productivity growth rate from −0.0 percent to 0.06 percent and the QD measure from −0.07 percent to 0.13 percent.

[15] As in Figure 2, the divergence between the two TFP estimates in Figure 3 is due to QD being flat in the late 1970s but having a higher growth rate than QT in the mid 1980s. Also, use of alternative deflators for the value of bank physical capital resulted in slightly lower productivity growth rates (a result similar to that obtained for the production-based measure of banking productivity—see previous footnote).

Figure 3

Cost Approach: Total Factor Productivity

(1977-87; 1977=100)



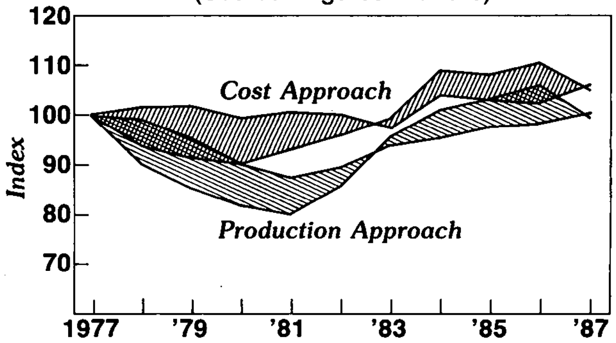
TFP(QT) = *Total factor productivity with transactions flow output*

TFP(QD) = *Total factor productivity with real balance stock output*

Figure 4

Comparison of Productivity Estimates Based on Production and Cost Growth Accounting Models

(Source: Figures 2 and 3)



derived from output and input quantities in Figure 2 are contrasted with those based on output cost and input prices in Figure 3. Results from the production approach suggest that productivity was mostly negative or zero over the period and therefore slightly lower than the cost approach, which yielded results showing zero to slightly positive productivity growth. In either case, the results show very low productivity growth, much lower than the annual 1.4 percent advance suggested in the BLS labor productivity series (Figure 2).

IV. ECONOMETRIC ESTIMATES OF BANKING PRODUCTIVITY

No studies, to our knowledge, have attempted to econometrically estimate TFP for U.S. banks.[16] Those U.S. studies that do exist have, instead, estimated only the effect of technical change. In a standard (translog) cost function context, $\ln C = f(\ln Q, \ln P_i, t)$, technical advance—indexed by time t—is expressed as $-\partial \ln C/\partial t$ while scale economies are $\partial \ln C/\partial \ln Q$. Total factor productivity is the combined effect of these two measures, adjusted for the change in output (dlnQ), or:

$$(5) \quad TFP = -\partial \ln C/\partial t + (1-\partial \ln C/\partial \ln Q)\, d\ln Q.$$

Estimates of technical change in banking have ranged from 0.96 percent a year over 1980-86 for a panel of 219 large banks (Hunter and Timme, forthcoming) to −0.90 percent over 1977-88 for a panel of 683 banks accounting for two-thirds of all bank assets (Humphrey, forthcoming).[17] In both of these studies, the scale economy estimate was so close to 1.00 that the scale adjustment to TFP in (5) has only a small effect (altering the annual values above to 1.05 and −1.01 percent, respectively). As seen in the table, the econometric estimates of banking TFP lie on either side of those from the growth accounting approach. Even so, all the estimates are relatively small, much less than one might have expected *a priori*.[18]

V. WHY WAS MEASURED BANKING PRODUCTIVITY SO LOW OVER THE LAST DECADE?

Cash Management and Deregulation: The Loss of Low-Cost Deposits

In the late 1970s, historically high interest rates greatly increased the use of cash management techniques by corporations. This meant large reductions

[16] Two studies do exist for other countries; one for Canada (Parsons, Gotlieb, and Denny, 1990) and another for Israel (Kim and Weiss, 1989).

[17] The −0.90 percent figure is from one of the preferred models estimated where bank physical capital is treated as a quasi-fixed input and a time-specific dummy variable is used (instead of a simple time trend) to reflect technical change. Two other studies of U.S. bank technical change exist (Hunter and Timme, 1986; Evanoff, Israilevich, and Merris, 1989) but these were concerned with only operating costs—not total costs—and are therefore not comparable with the analysis here.

[18] Indeed, the positive productivity growth rate from the Hunter and Timme (forthcoming) study can be turned into a small negative value when two deposit interest rates are specified in their model—one for core deposits, the other for purchased funds—rather than using the purchased funds rate for both as they did (see Humphrey, forthcoming, for details).

Box 2

Growth Accounting Measures of Banking Productivity[a]

Production Approach: Total Factor Productivity

Bank output (Q) is produced by combining the real value of capital (K), labor (L), demand deposits (D), small time and savings deposits (S), and purchased funds (F) inputs according to some production relation that changes in efficiency (A) over time: Q = A f(K, L, D, S, F). Expressed in terms of growth rates, the growth in total factor productivity ($\dot{A}/A$) is defined to be the difference between output growth and the expenditure share (w_i, i = K, L, D, S, F) weighted average of the growth in inputs:

Total Factor Productivity

$$(1)\quad \dot{A}/A = \dot{Q}/Q - w_K\dot{K}/K - w_L\dot{L}/L - w_D\dot{D}/D - w_S\dot{S}/S - w_F\dot{F}/F$$

where for X_i = Q, K, L, D, S, F:

$\dot{X}_i/X_i$ = an annual growth rate expressed as the index X_{it}/X_{it-1}, where t is time.

The use of expenditure share weights (w_i) presumes that the observed input prices—the rental price of capital, the wage rate, and the user cost of demand deposits, time and savings deposits, and purchased funds—equal the value marginal product of each input to the bank. When the w_i sum to 1.00, there is constant returns to scale.[b] The productivity measure (1) reflects total factor productivity (TFP) because the productivity effects of all inputs to the bank are being accounted for, along with returns to scale. While TFP is the most comprehensive measure of productivity, it is also the most difficult to compute because of the data required.

Multifactor and Single-Factor (Labor) Productivity

When more aggregative productivity measures are derived, such as for all manufacturing or all services, intermediate inputs are assumed to net out so only capital and labor inputs are used. The resulting measure is called multifactor productivity:

[a] This discussion is drawn from Hulten (1986).

[b] In the econometric approach to measuring productivity, the w_i are estimated statistically and need not sum to 1.00. In the growth accounting approach used here, the observed expenditure shares will sum to 1.00 by definition, imposing constant returns to scale. This restriction should only have a small effect on the results since numerous cross-section banking studies either support constant costs at the mean of all banks or are within 5 percentage points of it (so the cost elasticity of output ranges from slight economies of .95 to slight diseconomies of 1.05). See the surveys of Mester (1987), Clark (1988), and Humphrey (1990).

in idle demand deposit balances which did not pay explicit interest. The process is described and documented in Porter, Simpson, and Mauskopf (1979) and can be seen in Figure 5. Increased use of cash management techniques has emerged as the dominant explanation for the unexpectedly slow growth in the monetary aggregates during the 1970s. To compensate for the loss of demand deposits, banks came to rely more heavily on higher-cost purchased funds. Such a shift would have raised the real average cost per dollar of bank assets even if all input prices had remained constant. Since real average cost (corrected for input price changes) is the inverse of productivity, measured TFP would have fallen for this reason alone.

The negative cost effects from corporate cash management were continued with the banking deregulation of the early 1980s. Deregulation permitted noncorporate bank customers to switch from demand deposits to interest-earning Negotiable Order of Withdrawal (NOW) and Money Market Deposit Accounts (MMDAs). These new instruments inhibited the growth of demand deposits, shifting the

Multifactor Productivity

$$(2)\quad \dot{A}^*/A^* = \dot{Q}/Q - w_K\dot{K}/K - w_L\dot{L}/L$$

where $w_K + w_L = 1.00$.

The least comprehensive measure of productivity involves only the productivity of labor (LP) or output per unit of labor input: $LP = Q/L$. The growth in labor productivity is expressed as a reduced version of (1) or (2):

Labor Productivity

$$(3)\quad \dot{LP}/LP = \dot{Q}/Q - w_L\dot{L}/L.$$

Clearly, the growth of labor productivity in (3) will only equal the growth in TFP in (1) when labor is the only input (i.e., $w_L = 1.00$) or when the growth of other inputs are equal to that for labor (i.e., $\dot{L}/L = \dot{K}/K = \dot{D}/D = \dot{S}/S = \dot{F}/F$).

Cost Approach: Total Factor Productivity

All of the above equations showing productivity growth in terms of a production function have a corresponding cost function representation. That is, productivity can alternatively be expressed as the residual growth in average cost not accounted for by the growth in input prices over time. In simple terms, total factor productivity in a cost function context ($\dot{B}/B$) represents shifts in the average cost curve after controlling for changes in input prices:

Total Factor Productivity

$$(4)\quad \dot{B}/B = (\dot{C}/C - \dot{Q}/Q) - w_K\dot{PK}/PK - w_L\dot{PL}/PL - w_D\dot{PD}/PD - w_S\dot{PS}/PS - w_F\dot{PF}/PF$$

where:

$\dot{C}/C - \dot{Q}/Q$ = the growth rate of average cost, expressed as the growth in total cost less the growth in output; and

$\dot{PX}/PX$ = the growth rates of factor input prices and the user cost of funds, X = K, L, D, S, F.[c]

Under constant returns to scale, productivity growth using the production relationship in (1) equals minus one times the productivity growth from the cost relationship in (4) or $\dot{A}/A = -\dot{B}/B$.[d]

[c] The measurement of these variables is discussed in the Appendix.

[d] $\dot{A}/A$ is positive because increases in productivity in (1) increases output while $\dot{B}/B$ is negative as increases in productivity in (4) reduces cost.

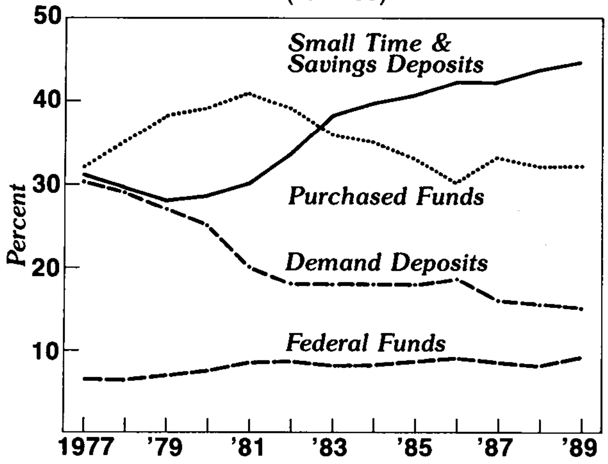
deposit expansion which did occur into interest-earning time and savings deposits (see Figure 5).[19]

Prior to deregulation, banks had substituted convenient branch offices, service personnel, and nonpriced services (e.g., free checking) for their inability to pay something close to a market rate on demand, savings, and small time deposits (Evanoff, 1988). Once deregulation removed interest rate ceilings and permitted consumer interest checking, banks quickly paid higher rates for the same funds. From a cost standpoint, banks subsequently found themselves to be "overbranched." The profitability of their deposit base fell from $61 billion in 1980, in constant 1988 dollars, to $4 billion in 1988 (Berger and Humphrey, forthcoming).

In effect, corporate cash management and deregulation removed banks' virtual monopoly control over zero-interest checking accounts and low-interest small

[19] While checks can be written on NOW and MMDA balances, they are not (legally speaking) available on demand and so have been classified with time and savings deposits in the data collected by regulatory authorities.

Figure 5
Percent Composition of Aggregate Bank Liabilities
(1977-89)



consumer time and savings deposits (as rate ceilings on these deposits were also removed).[20] Subsequent competition induced banks to shift from low-to higher-interest cost funds inputs without a fully offsetting reduction in factor inputs used to provide branch convenience and other low-priced deposit services. In addition, since the deposit services provided were largely unchanged as corporations conserved on idle balances and consumers shifted from one type of checking account to another, either measure of bank output used here would have been stable. With costs rising but output stable, costs per unit of measured output should rise, even when corrected for input price changes, lowering TFP.

In addition to cash management and deregulation, the inflation of the late 1970s and early 1980s also contributed to the rise in bank costs. During this inflationary period, some idle demand balances and low-cost time and savings deposits would have continued to shift to Money Market Mutual Funds (MMMFs) and been replaced by higher cost CDs sold by banks to the MMMFs. But in order to control operating costs, MMMFs restricted the number of checks written per month and specified high minimum amounts. Such limitations would likely have prevented any substantial disintermediation of demand deposits and thereby helped keep bank costs relatively low. Since over 80 percent of the deregulated bank balances were NOW and MMDA deposits (which experienced the largest rate increases following deregulation), it is clear that the great majority of the negative effects for banks seen during this period are due to deregulation, not inflation.

This analysis, we believe, explains why researchers have failed to observe much positive net technical change or productivity growth in banking during the last decade. Going beyond this explanation, part of the problem is also related to our inability to accurately capture all potentially important aspects of bank output. If branch convenience and the continued provision of underpriced deposit services are valued by users, then certainly some of the (now extra) costs incurred by banks in providing "unnecessarily" high levels of these services after deregulation have served to increase the quality of bank output. If one adopts this view, then what appears to be a productivity decrease may instead be the result of understating output growth as benefits received by bank depositors rose relative to their pre-deregulation level.

An analogous situation occurred in the electric utility industry during the 1970s. Expensive pollution control restrictions were mandated for electric utilities and, although these costs were largely made up by rate increases, measured output of this industry—kilowatt-hours—did not rise commensurately. As a result, measured total factor productivity was seen to fall (e.g., Gallop and Roberts, 1983). But if cleaner air resulted, then the quality of this industry's output actually rose but will not be captured in the output measure used. It is argued here that the same sort of thing occurred in banking.

Market-Share Reasons for Not Reducing Branch Convenience as Interest Costs Rose

It is easy to argue that the cost effect of deregulation could have been minimized if all banks had pared their branch operations more rapidly and to a greater degree. As it was, the real deposit/branch ratio was still falling until 1982, when it reached a minimum of around $28 million in core deposits per branch office. This meant that banks were still effectively building branches more rapidly than its customer base was expanding, increasing convenience (and operating costs) in the process. While the employee/branch ratio was more or less falling continually over this period, only after 1982 did the real deposit/branch ratio start to rise, reaching around $36 million in 1988.

Seemingly, market share considerations inhibited a more rapid and comprehensive reduction in bank

[20] Another aspect of deregulation was that thrift institutions obtained the ability to offer checkable deposits. This increased competition and contributed to the reduction in banks' monopoly power over this low-cost product.

operating costs as interest expenses from deregulation rose. Since choice of a bank by a depositor is largely based on convenience (according to industry surveys), a dramatic and profitable reduction in one bank's branching network would serve also to expand market share and profits at competing banks that retained their branch networks. In the end, both sets of banks would have experienced higher profit *rates* in the short run, but market shares and profit *levels* would have been redistributed away from those banks that cut their branch networks the most. Thus most banks seemingly chose to sacrifice short-term profits in order to maintain market share and hoped that long-term profit would follow as deposit growth continued to exceed the establishment of new branches.

Outlook for the Future

The outlook is not very bright. First, the wave of interstate mergers that have occurred already, along with those expected during the 1990s (when many states will eliminate their existing out-of-state merger barriers), bring with them costly "one-time" expenditures to integrate back office operations and standardize the banking products offered. While these expenditures will permit some cost reductions to be realized, they will also add considerable software and equipment expenses.

Second, the problem of excess banking capacity, as evidenced by too many branches, cannot easily be solved as long as failed or failing banks and thrifts continue to be purchased by institutions with the bulk of their own branch network typically outside of the purchased bank's deposit market area. Rarely do regulators simply close a failed bank's branches, and rarely do banks in the same market area purchase branches simply to close them. Instead, a failed bank's branch network is typically sold to an institution outside the market area and the buyer typically keeps most of the branches open, perpetuating the oversupply problem.

If the antitrust market concentration restrictions on bank mergers were considerably relaxed, then costs associated with overlapping branch networks would fall. Such cost reductions result when large competitors in the same deposit market area are encouraged to acquire each other and close excess branch offices (e.g., as occurred with Crocker and Wells Fargo in California). While market concentration would rise, it is not clear that increased concentration would or has led to much uncompetitive behavior in the form of reduced price competition and increased profits. Indeed, recent research indicates that low costs are the dominant explanation for higher bank profits in concentrated markets (Timme and Yang, 1990), not concentration itself as has long been asserted. Overall, given the two problems just outlined, it is hard to be optimistic about the future of productivity in banking. The most likely outcome is continued slow growth until the industry is able to shrink itself sufficiently through greater reductions in operating costs per dollar of deposits or assets. Thus future productivity growth will more likely stem from reducing current excess costs than from further technological progress.

VI. Summary

Measured productivity in banking over the last decade has been growing at a very low rate. Using aggregate data over 1977-87, it is estimated that total factor productivity growth has only been between −0.07 to 0.60 percent a year.[21] These estimates are based on a nonparametric growth accounting approach using first a production function and second a cost function. These results were robust to a number of influences (three different deflators for deriving the real value of bank physical capital and two different labor employment series). Importantly, these results are also robust to using two different indicators of banking output: one a flow measure of deposit and loan transactions and the other a stock measure of the real value of deposits and loan balances.

The primary explanation for the low productivity growth experienced has been the shift in zero-interest cost corporate and some consumer demand deposits to purchased funds in the 1970s (a result of improved corporate cash management techniques, higher interest rates, and the rise of Money Market Mutual Funds), plus a later shift of consumer demand deposits to interest-earning and checkable time and savings deposits in the 1980s (a result of banking deregulation which removed interest rate ceilings on time and savings and established new interest-earning checking accounts at both banks and thrifts). These developments significantly raised the cost of bank loanable funds. However, banks did not fully offset these higher costs by lowering operating expenses, reducing branch and service convenience, to compensate for the higher interest being paid. It is argued that market share considerations limited this response.

[21] Similarly low positive to low negative annual rates of productivity growth have also been found over a longer period, 1967-87 (Humphrey, 1991).

The outlook for the future is not bright. What is necessary is a substantial reduction in operating costs, since banking no longer has a virtual monopoly over zero-interest checking accounts and low-interest small consumer time and savings deposits. Future bank mergers, while reducing costs in some instances, will also lead to expensive "one-time" expenditures to integrate back office operations and standardize banking products. And bank failures, rather than removing excess branch office capacity as would occur in other industries, have tended to perpetuate the overcapacity conditions that have led to higher costs. Increases in banking productivity, when they come, are more likely to result from reductions in current operating costs and a rationalization of overlapping branch networks than from further technological progress.

APPENDIX

Availability of Data and Measurement of Banking Output and Price Indexes

Data Availability

Aggregate data on the number of deposit accounts from the FDIC are only available for two years over the past ten, while no aggregate data are available on the number of (new plus outstanding) loan accounts. While numbers of deposit and loan accounts are reported in the Federal Reserve's annual *Functional Cost Analysis* survey, the data cannot be used in a time-series analysis. First, the sampled banks change by upwards to 15 to 20 percent each year so that a consistent time series covering the same set of banks is not available. Second, the very largest banks, those that service the largest number of such accounts and experience the greatest rate of growth, are not included in the survey.

Indexes of Bank Output

The transactions flow index of banking output (QT) was developed by the Bureau of Labor Statistics (BLS, 1989). This index measures demand deposit output by the number of checks and electronic funds transfers processed, which reflects the debiting and crediting of demand deposit accounts as well as the payment processing and accounting activities associated with these activities. Similarly, savings and small denomination time deposit output is captured by measuring deposit and withdrawal activity in these accounts. Loan output is represented by the number of new real estate loans, consumer installment and credit card loans, and commercial, industrial, and agricultural loans made during the year. Lastly, trust and fiduciary activities are assumed to be proportional to the number of trust accounts serviced. Investment activities are treated as an intermediate good and netted out, since their variation has historically been associated with secondary reserves (where securities are sold to fund higher-than-expected loan demand or deposit withdrawal activity and vice versa). In any event, investment activities, plus the provision of safe deposit boxes, investment advice, and insurance, account for only a little more than 4 percent of bank employment, and their omission is not believed (by the BLS) to have a significant effect on the variation in measured output. Employment shares were used to weight these separate transaction flows into a single index of banking output.

The alternative index of the real value of deposit and loan account balances (QD) was developed by the author. It represents a cost-share weighted average of the dollar value of five deposit (demand deposits, small time and savings deposits) and loan categories (real estate loans, consumer installment and credit card loans, and commercial, industrial, and agricultural loans) from aggregate *Call Report* data. The cost-share weights are from the annual *Functional Cost Analysis* surveys for banks with more than $200 million in deposits. Nominal values of these five output categories were deflated by the GNP deflator to approximate real values.

Total Cost of Output and Input Prices

Total cost is from the *Call Report* and excludes double counting at the aggregate level by deleting the cost of purchased federal funds (see text). The price of capital is a bank-weighted average of the new contract cost per square foot of bank and office building space for nine regions of the United States reported in F.W. Dodge, *Construction Potentials Bulletin* (various years). Other capital price deflators were also used and their effects are noted in the text (footnote 14). The real value of bank physical capital used is book value deflated by the capital price index. The price of labor is total expenditure on labor divided by the number of full-time equivalent workers (both from the *Call Report*). The prices per dollar of each of the three funds categories are in terms of user costs, composed of the interest rate paid (i), the per dollar reserve requirement (RR), and the per dollar service charge income (SC). Following Hancock (1986), but neglecting FDIC deposit insurance costs, user costs (UC) are in general $UC = (i + r_{FF}\,RR - SC)/(1 + r_{FF})$, where r_{FF} is the rate on federal funds, a market rate. The denominator adjusts for the fact that the numerator costs are only fully realized at the end of a one-year period, rather than at the beginning. RR and SC are small for time and savings deposits and are difficult to separate out from those on demand deposits, for which i is zero. With these considerations in mind, our user costs are: $UC_D = (r_{FF}\,RR - SC)/(1 + r_{FF})$; $UC_S = i_S/(1 + r_{FF})$; and $UC_F = i_F/(1 + r_{FF})$. In implementation, total costs and the two factor input prices were deflated by the GNP deflator to reflect real values. User costs are already in real terms (see Hancock, 1986).

REFERENCES

Baily, Martin N., and Robert J. Gordon, "The Productivity Slowdown, Measurement Issues, and the Explosion of Computer Power," in William Brainard and George Perry (Editors), *Brookings Papers on Economic Activity* 2 (1988), The Brookings Institution, Washington, DC: 347-420.

Benston, George, and Clifford Smith, Jr., "A Transactions Cost Approach to the Theory of Financial Intermediation," *Journal of Finance* 31 (May 1976): 215-31.

Berger, Allen N., and David B. Humphrey, "Measurement and Efficiency Issues in Commercial Banking," in Zvi Griliches (Editor), *Output Measurement in the Services Sector*, University of Chicago Press (forthcoming).

Board of Governors of the Federal Reserve System, *Functional Cost Analysis*, National Average Report, Commercial Banks, Washington, DC (various years).

——————, *Consolidated Report of Condition and Income*, Washington, DC (various years).

Clark, Jeffery, "Economies of Scale and Scope at Depository Financial Institutions: A Review of the Literature," Federal Reserve Bank of Kansas City *Economic Review* 73 (September/October 1988): 16-33.

Cullison, William E., "The U.S. Productivity Slowdown: What the Experts Say," Federal Reserve Bank of Richmond *Economic Review* 75 (July/August 1989): 10-21.

Evanoff, Douglas D., "Branch Banking and Service Accessibility," *Journal of Money, Credit and Banking* 20 (May 1988): 191-202.

Evanoff, Douglas D., Philip R. Israilevich, and Randall C. Merris, "Technical Change, Regulation, and Economies of Scale for Large Commercial Banks: An Application of a Modified Version of Shephard's Lemma," Working Paper, Federal Reserve Bank of Chicago, Chicago, IL (June 1989).

F.W. Dodge Division, *Dodge Construction Potentials Bulletin*, Summary of Construction Contracts for New Addition and Major Alteration Projects, McGraw Hill, New York (various years).

Fixler, Dennis J., and Kimberly D. Zieschang, "User Costs, Shadow Prices, and the Real Output of Banks," in Zvi Griliches (Editor), *Output Measurement in the Services Sector*, University of Chicago Press (forthcoming).

Gallop, Frank M., and Mark J. Roberts, "Environmental Regulations and Productivity Growth: The Case of Fossil-Fueled Electric Power Generation," *Journal of Political Economy* 91 (August 1983): 654-74.

Hancock, Diana, "A Model of the Financial Firm with Imperfect Asset and Deposit Elasticities," *Journal of Banking and Finance* 10 (March 1986): 37-54.

Hulten, Charles R., "Productivity Change, Capacity Utilization, and the Sources of Efficiency Growth," *Journal of Econometrics* 33 (October/November 1986): 31-50.

Hunter, William C., and Stephen G. Timme, "Technical Change, Organizational Form, and the Structure of Bank Productivity," *Journal of Money, Credit and Banking* 18 (May 1986): 152-66.

Hunter, William C., and Stephen G. Timme, "Technological Change and Production Economies in Large U.S. Commercial Banking," *Journal of Business* (forthcoming).

Humphrey, David B., "Cost and Technical Change: Effects of Bank Deregulation," *Journal of Productivity Analysis* (forthcoming).

——————, "Flow Versus Stock Indicators of Banking Output: Effects on Productivity and Scale Economy Measurement," Working Paper, Federal Reserve Bank of Richmond, Richmond, VA (May 1991).

——————, "Why Do Estimates of Bank Scale Economies Differ?," Federal Reserve Bank of Richmond *Economic Review* 76 (September/October 1990): 38-50.

Kim, Moshe, and Jacob Weiss, "Total Factor Productivity Growth in Banking: The Israeli Banking Sector 1979-1982," *Journal of Productivity Analysis* 1 (1989): 139-53.

Mamalakis, Markos J., "The Treatment of Interest and Financial Intermediaries in the National Account: The Old 'Bundle' Versus the New 'Unbundle' Approach," *Review of Income and Wealth* 33 (June 1987): 169-92.

Mester, Loretta, J., "Efficient Production of Financial Services: Scale and Scope Economies," Federal Reserve Bank of Philadelphia *Economic Review* 73 (January/February 1987): 15-25.

Parsons, Darrell, Calvin Gotlieb, and Michael Denny, "Productivity and Computers in Canadian Banking," Working Paper, Department of Economics, University of Toronto, Canada (June 1990).

Porter, Richard, Thomas Simpson, and Eileen Mauskopf, "Financial Innovation and the Monetary Aggregates," *Brookings Papers on Economic Activity* 1 (1979), The Brookings Institution, Washington, DC: 213-29.

Sealey, Calvin, and James Lindley, "Inputs, Outputs, and a Theory of Production and Cost at Depository Financial Institutions," *Journal of Finance* 32 (September 1977): 1251-66.

Timme, Stephen G., and Won K. Yang, "On the Use of a Direct Measure of Efficiency in Testing Structure-Performance Relationships," Working Paper, Department of Finance, Georgia State University, Atlanta, GA (September 1990).

U.S. Department of Labor, Bureau of Labor Statistics, *Productivity Measures for Selected Industries and Government Services.* Bulletin 2322 (February, 1989): 170.

Wykoff, Frank C., "Commercial Banking Productivity Growth: Evidence from Large Bank Balance Sheets," Working Paper, Department of Economics, Pomona College, Claremont, CA (January 1991).

Survey Evidence of Tighter Credit Conditions: What Does It Mean?

Stacey L. Schreft and Raymond E. Owens[*]

Since early 1990, the results of the Federal Reserve Board's Senior Loan Officer Opinion Survey on Bank Lending Practices have been cited frequently as an indicator of general credit availability. Results from the Board's survey suggest that a considerable share of respondent banks were tightening their lending standards during 1990 and early 1991. How should these results be interpreted? This article attempts to answer this question by addressing the nature of the survey, examining the recent responses more closely and comparing recent results to past results.

A Brief History and Description of the Senior Loan Officer Survey

The Federal Reserve Board (hereafter, Board) first began conducting its Senior Loan Officer Opinion Survey in late 1964.[1] The survey was considered experimental until 1967, when it was made official and the Board began releasing its results to the public. Neither the survey's sample nor its format was changed from 1967 through 1977. Over this period, a sample of at least 121 banks from among those already participating in the Board's Survey of Terms of Bank Lending completed a written questionnaire each quarter. These respondents represented banks operating in the national business loan market, which accounted for 60 percent of business loans outstanding at all commercial banks.

The survey is qualitative rather than quantitative, focusing on loan officers' judgments about recent changes in their banks' non-price lending practices. Multiple- or dichotomous-choice questions are asked; that is, respondents must select a response from a list provided. From 1967 through 1977, the survey contained a consistent set of 22 questions, some of which were designed to identify whether banks' non-price lending policies (e.g., their standards of creditworthiness) were, on net, tighter, easier or unchanged from three months earlier. The Board reasoned that banks first responded to changes in the cost and availability of loanable funds by changing non-price lending terms and conditions of lending; only later would they adjust their interest rates. Therefore, information on changes in bank non-price lending policies would help explain the banking industry's response to monetary policy actions.[2]

The Board has revised the survey's format several times since 1977.[3] In February 1978, it changed several questions to capture more information on bank interest rate policies and on the willingness to make loans of different maturities. In May 1981, the sample was cut to 60 large U.S. commercial banks, generally the largest banks in their Federal Reserve districts.[4] Also at that time, the Board stopped conducting the survey through written questionnaires; instead, Federal Reserve Bank officers familiar with bank lending practices began conducting the survey through telephone interviews with senior loan officers at sample banks. In addition, the Board reduced the set of common questions from 22 to 6, dropping the questions on willingness to make term business loans. Allowance was made for the inclusion of questions on timely issues.[5] Since 1984, the survey format has been even more variable, with the number and type of questions usually changing from one survey to the next; even the number of surveys may vary

[*] We would like to thank Marc Morris for critical research assistance. Dan Bechter, Thomas Brady, Tim Cook, Bill Cullison, Tom Humphrey, John Scott and John Walter provided valuable advice and information. The views expressed in this paper are the authors' and do not necessarily reflect the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] From 1964 through 1977 the survey was called the Quarterly Survey of Changes in Bank Lending.

[2] See "Quarterly Survey of Changes in Bank Lending" (April 1968), pp. 362-63, and Taylor (1990).

[3] See Davis and Boltz (1978), Trepeta (1981) and Taylor (1990).

[4] In August 1990, 18 U.S. branches and agencies of foreign banks were added to the sample. See Brady (1990).

[5] Over the years, questions have appeared on subjects like the pricing of loan commitments, the use of standby letters of credit, the financial deterioration of business loan customers, the effect of money market deposit accounts on bank lending practices and home mortgage activity.

from year to year. Questions on standards of creditworthiness for business loans were not included from 1984 through early 1990.

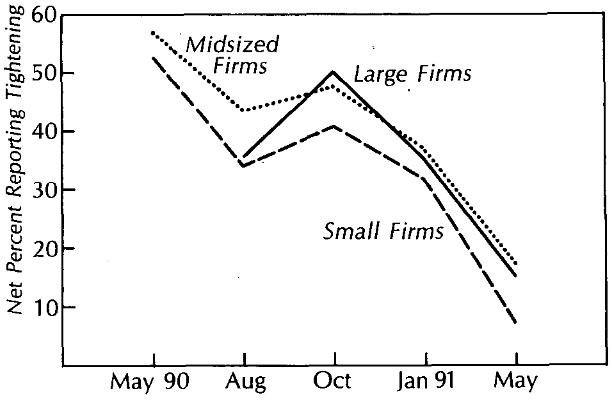
Recent Survey Results

In May of 1990, the Board reintroduced questions on business lending standards. Respondents were asked the following multiple-choice question: "Since late last year, how have your bank's credit standards for approving loan applications from C&I [commercial and industrial] loan customers changed for middle market firms and for small businesses?" Respondents could answer that their banks' credit standards had "tightened considerably," "tightened somewhat," been "basically unchanged," "eased somewhat" or "eased considerably." Changes in the enforcement of standards were to be reported as a change in standards.

The question remained in subsequent surveys, but the wording varied. In August and October of 1990 and January and May of 1991 the survey asked, "In the last three months, how have your bank's credit standards for approving applications for C&I loans or credit lines—other than those to be used to finance mergers and acquisitions—from large corporate, middle market and small business customers changed?"

Chart 1 shows the results from the May 1990 through May 1991 surveys, which have received considerable media attention.[6] It depicts the difference between the number of respondents reporting "tightened considerably" or "tightened somewhat" and those reporting "eased considerably" or "eased somewhat," as a percentage of all respondents. Hence, the larger the difference, the greater the net tightening of credit standards according to the survey results. On net, over 50 percent of respondents tightened standards for firms of all sizes during the first third of 1990, based on the May 1990 survey. Only one lender reported easing. The August survey showed over, 33 percent tightening further on loans to firms of all sizes; by October, at least 40 percent reported further tightening. At most 37 percent reported having tightened again on the January 1991 survey, while 17 percent did so on the May survey. No banks reported easing on the August, October or January surveys.

Chart 1
Changes in Bank Standards of Creditworthiness
(% Tightening Standards – % Easing)



Source: Federal Reserve Board Senior Loan Officer Opinion Survey.

Survey Results from Earlier Periods

How should the recent survey results be evaluated? Are the results more extreme than those found typically? Do they resemble results from surveys taken during past recessions or periods of comparatively slow credit growth? Answers to these questions can be gleaned from responses to similar questions asked in earlier surveys.

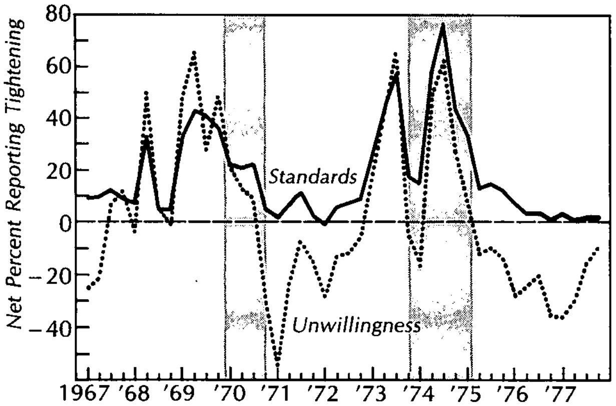
1967-77 Since the Senior Loan Officer Opinion Survey was initiated, the 1967-77 period has been the only extended period during which consistent questions about standards for and willingness to make business loans were asked. Chart 2 summarizes the responses to these two questions, neither of which is identical in wording to those asked recently. The solid line represents the responses of loan officers when asked how their banks had changed their "standards of creditworthiness for loans to nonfinancial businesses." Possible answers were "much firmer policy," "moderately firmer policy," "policy essentially unchanged," "moderately easier policy" and "much easier policy." As in Chart 1, the line depicts the difference between the number of respondents reporting "much firmer policy" or "moderately firmer policy" and those reporting "moderately easier policy" or "much easier policy," as a percentage of all

[6] Results are shown only for the 60 U.S. banks in the survey sample, not the branches and agencies of foreign banks. It is worth noting that the responses used to calculate the net percentages of respondents tightening lending standards or less willing to lend are not weighted by the asset size of the respondent banks. Thus, if the respondents reporting tighter lending standards generally have lower asset levels than those reporting easing, true or asset-weighted credit standards may have eased even though the survey might show more respondents tightening than easing. In practice, the fact that results are not weighted by asset levels has only been a problem to date for the period 1978-83. During that period, there were usually some respondents reporting tightening and some easing.

Chart 2

Standards and Unwillingness to Lend

Measures of Tightening of Lending Practices: 1967-1977



Note: Surveys were conducted in February, May, August and November of each year. The chart begins with data from February 1967.

Source: Federal Reserve Board Senior Loan Officer Opinion Survey.

respondents. An average of 18 percent more respondents reported firmer standards than reported easier ones over the 1967-77 period.[7]

The dotted line in Chart 2 shows loan officers' responses when asked how their banks' "willingness to make term loans to businesses" had changed. Officers chose from five responses ranging from "considerably less willing" to "considerably more willing." The line shows the net *un*willingness to lend: the difference between the number of respondents *less* willing and those *more* willing, as a percentage of all respondents. That is, the greater the difference, the less willing banks are to lend. On average, 2 percent more respondents reported being less willing than reported being more willing to lend.

Three general observations can be made from Chart 2. First, changes in willingness to lend and changes in net credit standards generally move together; in fact, the correlation between the two series is 0.88. That is, when banks are less willing to lend, they tighten credit standards.

Second, the chart indicates a more generalized tightening of standards and decreased willingness to lend before and during recessions (the shaded time periods). For example, consider the December 1969 to November 1970 recession. Both series peaked in May 1969, with 43 percent of all respondents indicating firmer standards of creditworthiness and 65 percent reporting decreased willingness to lend. In contrast, for the last three months of the recession banks firming credit standards outweighed those easing by only 5 percent; likewise, those more willing to lend dominated those less willing by 28 percent. For 1969—a year during which there was much speculation about whether a credit crunch was in progress—an average of 38 percent reported tighter lending standards, while an excess of 47 percent reported decreased willingness to lend.

The survey yielded similar results for the November 1973 through March 1975 recession. Both series peaked in August 1973 with over 57 percent of respondents on net reporting firmer standards and decreased willingness to lend. In 1973, as in 1969, on average the net percentage tightening was 38 while the net percentage reporting decreased willingness to lend was 30. Both series declined for November 1973 and February 1974 and then began rising again, reaching new peaks in August 1974. Results for the end of the downturn, as captured by the May 1975 survey, showed that a below-average percentage of respondents had somewhat firmer standards and a decreased willingness to lend.

A third observation from Chart 2 is that *respondents almost never reported a net easing of standards* on business loans.[8] During expansions, standards tightened less dramatically than during recessions (i.e., relatively fewer banks reported further tightening), but the number of respondents tightening continued to outweigh the number easing. We discuss this remarkable aspect of the survey results below.

1978-83 By 1978 the Board had evidence that the role of the prime rate was changing.[9] Consequently, in revising the survey, the questions on business lending standards were rewritten to reflect that evidence. From 1978 through 1983, loan officers surveyed were asked about changes, compared with three months earlier, in their institutions' "standards of creditworthiness to qualify for the prime rate" and their standards "to qualify for a spread above prime." Possible responses were "much firmer," "moderately firmer," "essentially unchanged," "moderately easier" and "much easier." For a shorter period—1978 through February 1981—respondents were also asked about changes in their willingness to make

[7] Of banks *not* reporting a tightening of standards, the vast majority reported lending standards essentially unchanged from 1967 to 1977 and from 1978 to 1983.

[8] The February 1972 survey is an exception; one more respondent (0.80 percent) reportedly eased than tightened that quarter.

[9] See Brady (November 1985).

fixed-rate short-term (with maturities of less than one year) loans and fixed-rate long-term (maturities of one year or longer) loans. The five possible responses ranged from "considerately greater" to "much less." Responses to the two questions on lending standards were highly correlated, as were those on the two questions on willingness to lend.

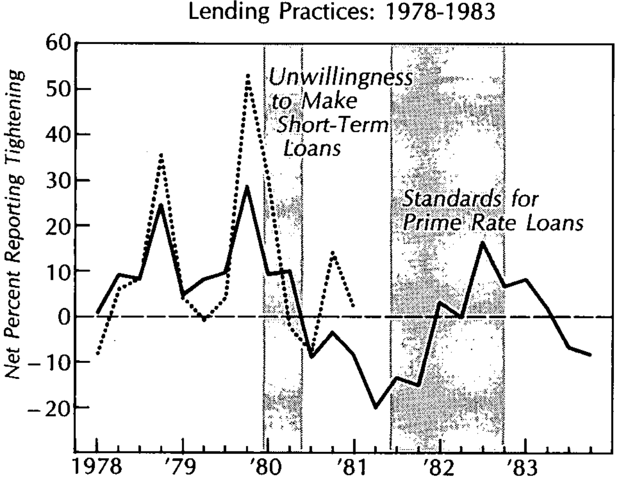
Chart 3 depicts reported changes in lending standards on prime rate loans and willingness to make fixed-rate, short-term loans. The results from the February 1978 through May 1980 surveys were similar to those from the 1967 through 1977 period. Specifically, a net tightening of standards was always reported, and changes in the willingness to lend are highly correlated with changes in lending standards. Moreover, the net tightening of standards reached a peak with the survey preceding the 1980 recession (the November 1979 survey). This peak of 29 percent is lower than the peaks preceding the two earlier recessions.

In contrast, the results for the August 1980 through November 1981 surveys deviated considerably from those for 1967 through mid-1980. For this period, respondents reported a ***net easing*** of lending standards. These results are particularly perplexing because they are the only evidence of a net easing over a 15-year period. The July 1981 through November 1982 recession is preceded by an easing of standards that "peaks" in May 1981, with 20 percent more respondents saying that they were easing policy, most of them doing so "moderately," than saying they were tightening. For the question (not shown in the chart) about changes in standards to qualify for a given spread ***above*** prime, the results are more extreme: 42 percent reported easing on net. Throughout the recession, a tightening of standards was reported on net by at most only 17 percent of respondents, approximately the average for the 1967-77 period.[10]

Chart 3

Standards and Unwillingness to Lend

Measures of Tightening of
Lending Practices: 1978-1983



Note: Surveys were conducted in February, May, August and November of each year. The chart begins with data from February 1978.

Source: Federal Reserve Board Senior Loan Officer Opinion Survey.

What explains these anomalous survey results? As Brady (1985) has documented, a weakening of the link between prime rates and market rates took place during the 1970s. Banks began pricing loans to large borrowers at market rates and, to a great extent, reserving the prime rate and prime-based rates for smaller and less creditworthy borrowers.[11] From mid-1980 through 1981, the prime rate was ***above*** the average loan rate (Chart 4). With the margin on ***prime rate*** loans comparatively high, lenders depended more on interest rates and less on standards of creditworthiness as a means of allocating credit. It is not surprising then that survey respondents reported an even more pronounced easing of standards on ***above-prime rate*** loans that had even higher rates relative to the average loan rate.

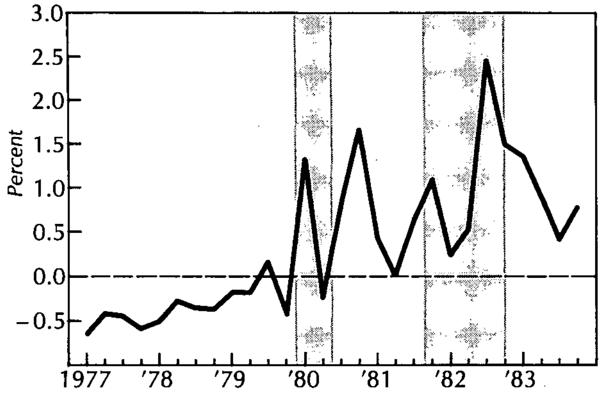
With the survey results for mid-1980 through 1981 accounted for, we conclude that the trends observed for the 1967-77 period continued to hold for 1978 through 1983. As stated above, no questions on the standards of creditworthiness for business loans appeared on the survey from 1984 until May 1990.

[10] The question on willingness to make fixed-rate short-term loans was not asked after February 1981, but its relationship to the standards question probably would have remained unchanged, given the high correlation between the two questions (a correlation of 0.76 from February 1978 through February 1981), had it been asked.

[11] Brady (November 1985, pp. 21-22) explains that interest rates (both market rates and the prime rate) were relatively stable until the mid-1960s. Thus, prime-based loan pricing, which was common during this period, resulted in relatively stable loan rates. The relationship between market rates and the prime rate began to change throughout the 1970s as market rates became more variable and U.S. branches of foreign banks, which priced loans off market rates, competed more actively in the U.S. commercial loan market. By about 1982, the practice of linking loan rates to market rates, which represented the marginal cost of funds, rather than to the prime, apparently a measure of the average cost of bank funds, was commonplace. As a measure of average costs, the prime changed more slowly in a volatile rate environment than did market rates. Thus, borrowers could obtain relatively stable interest rates with prime-based loans. Brady suggests that small borrowers may have preferred this stability.

Chart 4

Spread: Prime minus Weighted Average Short-Term C&I Loan Rate



Note: Quarterly data are shown beginning with the first quarter of 1977.
Source: Federal Reserve Board Quarterly Terms of Bank Lending.

Interpreting the Recent Results

Looking at survey results from an historical perspective shows that recent responses resemble those from the 1969 to 1970 and 1973 to 1974 recessions.[12] Specifically, for the years 1969 and 1973, 38 percent of respondents on net reported a further tightening of lending standards, more than double the percentage on average from 1967 through 1983. During 1990, at least 40 percent reported further tightening on average.[13] The 1991 survey results thus far (those for January through May) closely match those from the middle of both the 1969 to 1970 and 1973 to 1975 recessions. The May 1991 survey indicated net tightening by at most 17 percent, the average for the 1967 to 1983 period.[14]

It is also worth noting that from 1967 through 1983 respondents almost never reported a net *easing* of standards on business loans; in fact, net tightening was reported by an average of 17 percent of respondents.[15] This suggests that the survey responses might be biased. Why might bias arise? One possible reason stems from the incentive that regulated institutions have to report to their regulator a tightening of standards, especially when their reports are not made anonymously. This incentive would exist if respondent banks perceive a risk of closer regulatory scrutiny if they admit to having eased standards. During 1990, this risk might have been perceived as especially great, given reports that many bankers viewed regulators as being overzealous in their examination of loan portfolios.[16]

The persistent reports of tighter credit conditions over the history of the survey make the survey's *absolute* numerical results (that is, the net percentage of banks tightening) difficult to interpret. To some extent, however, the pattern of the reports of tightness across business cycles means that the survey's results are most meaningful when viewed *relative* to those from previous periods. Noting this, the recent results of a tightening of lending standards by a considerable share of respondents appear to be typical for an economy entering or in a recession.

[12] We cannot compare the recent results to those for the 1980 or 1981 to 1982 recessions because the survey during those periods asked about standards on prime rate and above-prime rate loans and thus are not comparable, as discussed above.

[13] Recall that the 1990 surveys asked about standards to large, middle-market and small firms. The average over the surveys conducted in 1990 is at least 40 percent for firms in each category.

[14] Each quarter since 1973, the National Federation of Independent Business has surveyed its membership about their borrowing experiences. Dunkelberg (1991) analyzes the results and finds that the net percent of members reporting credit being harder to get during 1990 and the first quarter of 1991 is *low* relative to that in 1974 and 1980.

[15] Remember that the survey results are essentially first differences: they report the change in lending standards over a three-month period, not how tight standards are at the survey date. Thus, because the results show banks continuously tightening their standards from 1967 through 1983, if we take the survey results literally, lending standards would have been unbelievably stringent by late 1983.

[16] Despite these reports, relatively few survey respondents cited regulatory pressures as the cause of their tightening of lending standards.

The Consumer Installment Loan Question

Only one item has appeared consistently on the Senior Loan Officer Opinion Survey: "Indicate your bank's willingness to make consumer installment loans now as opposed to three months ago" (as worded on the January 1991 survey). Possible responses were "much more," "somewhat more," "about unchanged," "somewhat less" and "much less." Chart 5 displays the difference between the number less willing and the number more willing, as a percentage of all respondents. Answers to this question exhibit the same patterns around recent business cycles as do the answers regarding willingness to make business loans. However, the 1980 results are extreme. On the May 1980 survey, those reporting being less willing to make consumer installment loans exceeded those indicating greater willingness by 57 percent, a record number and well above the −42 percent level recorded in the August 1980 survey. The May survey was conducted while selective credit controls were in place, and it asked lenders to compare their willingness to lend in May with that in February, before the control program began. One component of the controls was a 15 percent reserve requirement on all extensions of consumer credit over some base amount.[a] The controls were lifted in early July, and by August the economy had rebounded from its spring slump. Lenders were once again willing (and encouraged by policymakers) to lend.

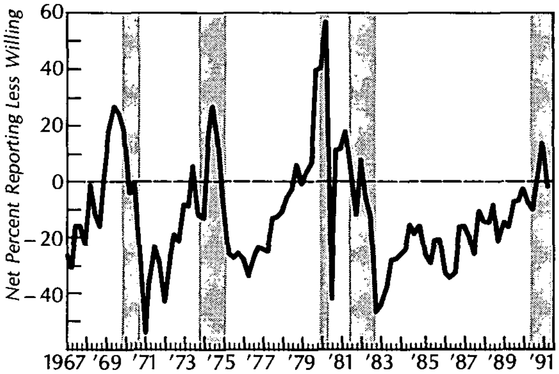
[a] Schreft (1990) examines the 1980 credit control program in depth.

Chart 5

Unwillingness to Make Consumer Loans

A Measure of Tighter Lending Practices

1967-1991



Note: Surveys were conducted in February, May, August and November of each year. The chart begins with data from February 1967.

Source: Federal Reserve Board Senior Loan Officer Opinion Survey.

References

Brady, Thomas F. Memo entitled "The August 1990 Senior Loan Officer Opinion Survey on Bank Lending Practices," Board of Governors of the Federal Reserve Board, August 1990.

——————. "The Role of the Prime Rate in the Pricing of Business Loans by Commercial Banks, 1977-84," Staff Study No. 146. Washington: Board of Governors of the Federal Reserve System, November 1985.

Davis, Patricia, and Paul Boltz. Memo to Mr. Lindsey on Senior Loan Officer Opinion Survey on Bank Lending Practices, Board of Governors of the Federal Reserve System, March 23, 1978.

Dunkelberg, William C. "The Credit Crunch—Myth Or Mistaken Monetary Policy?" National Federation of Independent Business, April 1991.

"Quarterly Survey of Changes in Bank Lending," *Federal Reserve Bulletin*, April 1968, pp. 362-63.

Schreft, Stacey L. "Credit Controls: 1980," Federal Reserve Bank of Richmond, *Economic Review*, Vol. 76, No. 6 (November/December 1990), pp. 25-55.

Taylor, Gail Ann. Memo to Board Committee on Research and Statistics on Proposal for Extension, with Revision, to the Senior Loan Officer Opinion Survey on Bank Lending Practices (FR 2018), Federal Reserve Bank of San Francisco, March 26, 1990.

Trepeta, Warren T. Memo to Mr. Simpson on Senior Loan Officer Opinion Survey on Bank Lending Practices, Board of Governors of the Federal Reserve System, June 19, 1981.