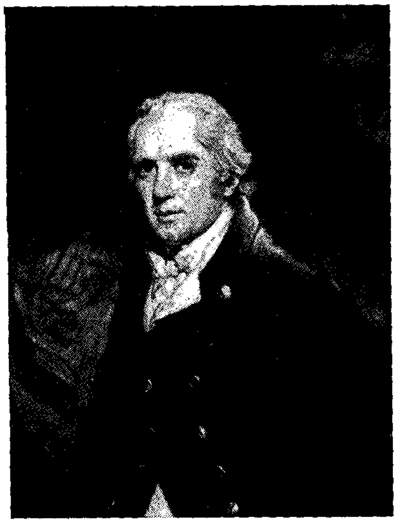
HENRY THORNTON: SEMINAL MONETARY THEORIST AND FATHER OF THE MODERN CENTRAL BANK

Robert L. Hetzel

1. Introduction

In 1802, Henry Thornton published the book *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain.*[1] On the basis of this work, Thornton deserves the title of "Father of Modern Central Banking." Thornton developed the idea of a central bank that could control the monetary base as a bookkeeping operation. Through control of the base, the central bank could control the money stock of the entire country. Finally, through control of the money stock, the central bank could control the price level. A key theme of *Paper Credit* is the importance of explicit acceptance by the central bank of its responsibility for determining the price level. Not until Keynes' *A Tract on Monetary Reform* is there again such a forceful statement of the concept of a modern central bank.



Reproduced by permission of Edward Arnold Ltd.
From *Marianne Thornton* by E. M. Forster.

In 1810, Thornton repeated these ideas in the *Bullion Report.* Although this report was written jointly by Horner, Huskisson, and Thornton, the analytical framework used is Thornton's. The *Bullion Report* is reviewed in the final sections of this article as a way of showing the use which Thornton made of the analytical apparatus he developed in *Paper Credit.*

Thornton analyzed the paper money standard that existed in Britain after suspension of the international gold standard in 1797. For this fiat money regime, Thornton constructed a general equilibrium model capable of explaining the relationship between the domestic price level and the exchange rate and capable of explaining movements in the exchange rate either as a real phenomenon or a monetary phenomenon. The chief operating variable of the Bank of England was the discount rate. Thornton developed an exposition of the quantity theory organized around the differing role of the interest rate in the supply and demand schedules for the money stock. As a condition for maintaining a stable monetary base and money stock, the supply schedule required the central bank to keep the discount rate in line with the economy's natural rate of interest either by rationing explicitly its discounts or by targeting a nominal variable like the exchange rate.

On the basis of the contributions in this book, Thornton deserves to be ranked among the foremost monetary theorists of all times. Only a small number of economists, however, are aware of his contributions. There are two reasons for this lack of recognition.

First, Thornton organized his economic analysis around the central proposition that with a noncommodity monetary standard based on the fiduciary issue of banks a central bank must assume explicit control over its own liabilities (the monetary base). This control is necessary in order to maintain the money stock and to maintain a well-defined price level. When the international gold standard became enshrined as monetary orthodoxy in the last half of the nineteenth century, the idea of a central bank

This paper was written on the occasion of the 75th birthday of Milton Friedman, July 31, 1987.

[1] Page references are to the Hayek edition, which includes *Paper Credit,* two speeches Thornton made before Parliament in May 1811, and a biographical sketch of Thornton by Hayek. Pages are indicated by (H, page -). The page citations in the paper (C, page -) are to the Cannan edition of the *Bullion Report.*

exercising explicit control over the monetary base became only a theoretical curiosum. Under the international gold standard, the balance of payments, not the behavior of the central bank, was supposed to determine the monetary base and the nominal quantity of money. As a consequence, Thornton's work was ignored by neoclassical economists.[2]

The second reason for the obscurity of Thornton's work is Thornton's own style of exposition. The ideas in *Paper Credit* are exposited according to the chronological order in which Thornton dealt with particular problems of policy, rather than being exposited in a way designed to elucidate the underlying analytical framework. More important, this underlying framework is nowhere succinctly presented, but is submerged in a great mass of institutional detail. In his review of *Paper Credit* in the *Edinburgh Review*, Francis Horner (1802, p. 29) states: "But the various discussions are so unskilfully arranged that they throw no light on each other, and we can never seize a full view of the plan. . . ." Later economists reviewing the monetary debates at the beginning of the nineteenth century turned to David Ricardo. Ricardo's quantity theory framework was only a caricature compared to Thornton's, but the clarity and forcefulness with which Ricardo exposited his framework made him, rather than Thornton, the more accessible author.

Thornton's work is discussed in some of the classic works in economics. Viner (1924) discovered Thornton and discussed his contribution to the theory of international trade. Later, Viner (1937) also discussed Thornton in the context of the bullionist-antibullionist controversy. Hayek (1931) was interested in Thornton because of the latter's concept of a modern central bank that can control the monetary base and the money stock. In his introduction to the reprint of *Paper Credit*, Hayek carefully lists the seminal ideas of Thornton. Another such list is in Hutchison (1968). Mints (1945) reviews Thornton's criticisms of the real bills principle. Schumpeter (1954) insightfully notes the relationship between Thornton's and Wicksell's views of credit creation. [See also Humphrey (1985).] Despite these discussions, there remains a need for an overview of the analytical framework employed by Thornton. This essay is motivated by the belief that the major reason for the current lack of appreciation of Thornton is the absence from the literature of a comprehensive overview of the general equilibrium model of the economy developed by Thornton.

Section 2 presents some of the historical background to Thornton's work. Section 3 explains Thornton's goal of extending the quantity theory to include not only specie, but also money created through credit extension. Section 4 discusses Thornton's theory of money demand. As background to Thornton's theory of money supply, Section 5 discusses the natural rate hypothesis built into Thornton's theory of aggregate supply and demand. Section 6 discusses Thornton's theory of money supply. Section 7 presents Thornton's criticisms of the real bills view. Section 8 contains Thornton's discussion of the monetary consequences of the international adjustment mechanism. Section 9, which begins the discussion of the *Bullion Report*, reviews the antibullionist views of the Governors of the Bank of England who testified before the Bullion Committee. It also presents the rebuttal of these views by the bullionists. Section 10 contains the recommendations of the Bullion Committee about the appropriate policy for the Bank of England. Section 11 contains a summary of the article, and Section 12 discusses the relevance of issues raised by Thornton for modern central banks.

2. Historical Background[3]

Hayek notes that "Since the contributions of Cantillon, Galiani, and Hume in the middle of the eighteenth century little progress had been made in monetary science. . . . And the treatment of money in the *Wealth of Nations*, which dominated opinion on these matters in the last quarter of the century, contains comparatively little of theoretical interest" (H, 37).[4] Toward the end of the eighteenth century, however, significant changes in institutional arrangements prompted an interest in issues of monetary policy. The number of country banks increased rapidly, and the Bank of England became the sole issuer of bank notes in London. In 1797 in testimony before Parliament, Francis Baring, in characterizing the Bank of England, first used the expression bank of dernier resort (last resort). The financial panic of 1793 and the ensuing increased demand for Bank of England notes encouraged reflection on the special role of that bank in the banking system. The war with France, which began in 1793,

[2] The work of the classical economists was known to the neoclassical economists primarily through the writings of Ricardo and through J. S. Mill's *Principles of Political Economy.* Mill mentions Thornton only in the context of a discussion of the origin of bills of exchange.

[3] The first two paragraphs and the last paragraph of this section draw on Hayek's introduction to *Paper Credit.*

[4] Humphrey (1981), however, argues that this view must be qualified by recognizing Smith as advocating the view now known as the monetary approach to the balance of payments.

over time created a situation that forced Britain off the gold standard. Gold was sent out of England due to British financial support of its continental army and allies. Also, the return of France to the gold standard under Napoleon increased the demand for gold. Finally, in 1797, fear of a French invasion precipitated a run on the gold reserves of the Bank of England. This run led the Bank of England to suspend redemption of its notes in gold.

At first, the experiment with a noncommodity money standard went well. There was little inflation or depreciation of the pound on the foreign exchanges. Gold flowed into England and the Bank of England replenished its reserves. The situation deteriorated beginning in 1800, however. Borrowing by the British government from the Bank increased. Domestic prices began to rise and the pound depreciated on the foreign exchanges. Because Napoleonic Europe was on the gold standard, the pound price of gold bullion measured the foreign exchange value of the pound. In 1800, the pound price of bullion rose to a value ten percent in excess of the old mint price under convertibility. The rise in the market price of gold over the old mint price prompted criticism of the Bank of England for having suspended convertibility.

This historical chronology explains the organization of *Paper Credit.* Thornton deals first with the appropriate response of the Bank of England to an internal drain that produces a financial panic. He views this problem not just in terms of bank runs, but also in terms of an increase in the precautionary demand for money. Thornton elaborates a sophisticated theory of the demand for money that first explains the way in which credit creation leads to money creation in a fractional reserve system and then relates the velocity of the various components of money to the difference between the market rate of interest and the own rate on the various components. Thornton defends the suspension of cash payments by the Bank of England in 1797 as necessary in order to prevent a contraction of the money stock and the associated adverse consequences for real economic activity.

In the last part of his book, Thornton considers the key dispute between the bullionists and antibullionists. He considers the dispute over whether the depreciation of the pound on the foreign exchanges that began after 1800 was caused by an adverse movement in the commodity terms of trade or currency overissue by the Bank of England. In order to consider this dispute, Thornton constructs an analytical apparatus general enough to explain both nominal and real movements of the exchange rate. In showing how an increase in the money stock can lead to a rise in the price level and a fall in the nominal exchange rate, Thornton elucidates the interaction between the central bank's discount rate, the economy's natural rate of interest, money creation, and the foreign exchange value of the pound. This discussion also contains an elaboration of a natural rate hypothesis to reconcile the short-run nonneutrality of money with long-run neutrality. Thornton uses his analytical apparatus to advance his central theme "that the restriction of the paper of the Bank of England is the means both of maintaining its own value and of maintaining the value, as well as of limiting the quantity, of all the paper in the country" (H, 225).

In *Paper Credit,* the main practical concern of Thornton had been disruption of economic activity from deflation, produced from maintenance of the international gold standard at a time of bank runs or a deterioration in the terms of trade. In the decade after the publication of *Paper Credit,* events caused his main concern to shift to inflation due to overissue by central banks. In one of his speeches before Parliament in 1811, Thornton says, "Indeed, in all parts of Europe, Hamburgh, Amsterdam, and Paris excepted, the principle of a standard seemed to have been lost; a suspension of cash payments had every where taken place; and [the] paper had been issued to excess, and had also been depreciated" (H, 342). Thornton refers in particular to the experiences of Sweden, Austria, and Portugal, as well as the earlier experience of Russia. As a member of Parliament's Committee on the Irish Currency, Thornton had a firsthand view of the overissue of the Bank of Ireland, which had suspended cash payments at the same time as the Bank of England. In England, the market price of bullion remained fairly close to its old parity price from 1804 through 1808. Beginning in 1809, however, the pound price of foreign exchange and bullion rose about 30 percent above the old parity. In 1801, the depreciation had been limited to 10 percent.

In 1810, against a background of rising prices and a falling price of the pound in the foreign exchange markets, Francis Horner moved before Parliament that a Select Committee be "appointed to enquire into the Cause of the High Price of Gold Bullion, and to take into consideration the State of the Circulating Medium and of the Exchanges between Great Britain and Foreign Parts." The resulting Bullion Committee Report was drafted by Horner, Huskisson, and Thornton. Its major recommendation was that Britain return to the gold standard in order "to enforce . . . a due Limitation of the Paper of the Bank of England, as well as of all the other Bank Paper of the Country. . . ." (C, Resolutions, 10) Thornton delivered two speeches, later issued as pamphlets, during the debate over the Bullion Committee Report. In these speeches, he repeats his model of credit and money stock determination, which turns on the difference between the Bank discount rate and the natural rate of interest. This time he supplemented his model with an explanation of the natural rate of interest as the sum of a real rate of interest and a liquidity premium dependent upon expected inflation.

3. Relationship between Credit Creation and Money Creation

Thornton extended the analytical apparatus of the quantity theory to money created through credit extension. In modern jargon, he extended the quantity theory to include not only outside money (the monetary base), but also inside money (the fiduciary issue of banks minus their reserves).

> Paper constitutes, it is true, an article on the credit side of the books of some men; but it forms an exactly equal item on the debit side of the books of others. It constitutes, therefore, on the whole, neither a debit or a credit. . . . The case of gold, on the other hand, differs from that of paper inasmuch as the possessor of gold takes credit for that which no man debits himself. (H, 79)

Thornton uses the term "paper credit" for inside money. The incentive for fiduciary issue, the issue of paper money, came from economizing on the real resource costs of a commodity money.

> When confidence rises to a certain height in a country, it occurs to some persons that profit may be obtained by issuing notes, which purport to be exchangeable for money; and which, through the known facility of thus exchanging them, may circulate in its stead; a part only of the money, of which the notes supply the place, being kept in store as a provision for the current payments. On the remainder interest is gained, and this interest constitutes the profit of the issuer. (H, 90)

Thornton was the first economist to assert that checking accounts formed part of the money stock.

> It is in substance the same thing whether a person deposits 100 pounds in money with the bank, taking no note, but obtaining a right to draw a draft on a banking account which is opened in his name, or whether he deposits the same 100 pounds and receives for it a bank note. (H, 134)

There were a few economists in the nineteenth century who viewed checking accounts as money, Torrens and Joplin, for example, and some economists in the banking school tradition. It was not until the 1920s, however, that economists working in the quantity theory tradition generally accepted these accounts as money. Unlike most other nineteenth century economists, Thornton was able to abstract from the legal distinctions distinguishing gold from fiduciary instruments embodying a claim to gold [Schumpeter (1954), 717]. He successfully integrated into his view of money all media of exchange based on credit creation.

In expanding the definition of money to instruments derived from credit extension, Thornton continually insists on the difference between the demand for money and the demand for credit.

> . . . it is by the amount not of the loans of the Bank of England, but of its paper . . . that we are to estimate the influence on the cost of commodities. (H, 271)

In applying the quantity theory to a monetary regime of paper money, Thornton begins with the distinction between relative prices and the price level. It is only with respect to the latter concept that the supply and demand analysis of the quantity theory is applicable.

> . . . the price at which the exchange (or sale) takes place depends on two facts; on the proportion between the supply of the particular commodity and the demand for it, which is one question; and on the proportion, also, between the state of the general supply of the general *circulating medium and that of the demand for it, which* is another. (H, 194)

4. The Demand for Money

According to Thornton, the demand for money includes both a transactions and a precautionary demand.

> The supply of bank notes which he chuses to reserve in his drawer is always estimated by the scale of his payments; or, to speak more correctly, by the probable amount of the fluctuations in his stock of notes, which fluctuations are proportionate, or nearly proportionate, to the scale of his payments. (H, 234-35)

> Now a high state of confidence contributes to make men provide less amply against contingencies. . . . When, on the contrary, a season of distrust arises, prudence suggests that the loss of interest arising from a detention of notes for a few additional days should not be regarded. (H, 96-97)

Thornton thought that increases in the precautionary demand for money acted to exacerbate financial panics. It is interesting to note in this respect that he criticizes the common notion of hoarding as an unsophisticated expression of the precautionary demand for money.

> When a season of extraordinary alarm arises, and the money of the country in some measure disappears, the guineas, it is commonly said, are hoarded. In a certain degree this assertion may be literally true. But the scarcity of gold probably results chiefly from the circumstance of a considerable variety of persons, country bankers, shopkeepers, and others, augmenting, some in a smaller and some in a more ample measure, that supply which it had been customary to keep by them. . . . It is thus that a more slow circulation of guineas is occasioned; and the slower the circulation, the greater the quantity wanted in order to effect the same number of money payments. (H, 99-100)

Thornton argues that the demand for components of the money stock varies inversely with difference between the market rate of interest and the own rate on the particular component.

> Bills, however, and especially those which are drawn for large sums, may be considered as in general circulating more slowly than either gold or bank notes. . . . Bank notes, though they yield an interest to the issuer, afford none to the man who detains them in his possession; they are to him as unproductive as guineas. The possessor of a bank note, therefore, makes haste to part with it. The possessor of a bill of exchange possesses, on the contrary, that which is always growing more valuable. . . . such part of the circulating medium as yields an interest to the holder will effect much fewer payments, in proportion to its amount, than the part which yields to the holder no interest. (H, 92 and 94)

5. Aggregate Supply and Demand

The Transitory Nonneutrality of Money Before discussing Thornton's money supply function, it is necessary to discuss his aggregate supply of output function. The long-run neutrality of money incorporated into this latter function endows Thornton's general model with a natural rate of interest. As discussed below, money supply is a function of the difference between the Bank rate and the natural rate.

In *Paper Credit,* Thornton emphasizes the economic disruption of deflation. A major theme in the book is a defense of the suspension of convertibility by the Bank of England in 1797. As noted in Section 8, Thornton believed that the terms of trade had changed adversely for Britain. Maintaining convertibility would, therefore, have required a deflation of the domestic British price level. Suspension of convertibility allowed the pound to depreciate on the foreign exchanges without this deflation. Thornton (H, 117-18 and 152) admonishes against deflation as a corrective to an adverse balance of trade.[5]

Thornton refers briefly to two reasons why a change in the money stock affects real economic activity. One reason is that wage rates do not respond to changes in prices perceived to be temporary.

> 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. [Italics in original] (H, 119)

Thornton also suggests that individuals confound changes in relative prices with changes in the price level.

> 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. While paper is encreasing, and articles continue rising, mercantile speculations appear more than ordinarily profitable. The trader, for example, who sells his commodity in three months after he purchased it, obtains an extra gain, which is equal to such advance in the general price of things as the new paper has caused during the three months in question: he confounds this gain with the other profits of his commerce; and is induced, by the apparent success of his undertakings, to pursue them with more than the usual spirit. (H, 237-38)

> . . . nations in general were usually insensible at first to the declension of the value of their circulating medium. They were accustomed to experience fluctuation of [their] exchange [rate], and they naturally referred, at first, even a serious depreciation of their paper, to the same commercial causes which they were in the habit of contemplating. . . . It was reasonable to suppose that men should generally mistake in this respect. We naturally imagine that the spot on which we ourselves stand is fixed and that the things around us move. The man who is in a boat seems to see the shore departing from him. . . . In consequence of a similar prejudice, we assume that the currency which is in all our hands, and with which we ourselves are, as it were, identified, is fixed, . . . whereas in truth, it is the currency of each nation that moves. (H, 340)

Forced Saving Thornton did believe that the seigniorage from money creation could redistribute income in such a way as to increase the capital stock and, thereby, to increase permanently the level of economic activity. He argues that ". . . borrowers, in consequence of that artificial state of things which is produced by the law against usury, obtain their loans too cheap" (H, 255) from the Bank of England.[6]

[5] An additional reason why Thornton favored suspension of the gold standard was that suspension allowed the export of gold coin. This export of gold coin mitigated the deterioration of the British terms of trade (H, 153).

[6] Thornton claims that only the Bank of England was effectively bound by the usury law. A borrower from a regular bank "bestowed the benefit of his running cash," that is, maintained compensating balances to raise the effective loan rate to the market clearing rate. A borrower in the money market "gave to a broker a small percentage on every bill," that is, paid points (H, 335).

The seigniorage from money creation goes to those able to borrow at a below market rate from the Bank. Income is redistributed to these individuals and away from holders of existing cash balances and wage earners whose wages are slow to adjust to inflation.

> The proprietors of the new paper will become greater encouragers of industry than before; the owners of the old paper, being able to command less property, will have less power of employing labour. . . . (H, 237) It must be also admitted that, provided we assume an excessive issue of paper to lift up, as it may for a time, the cost of goods though not the price of labour, some augmentation of stock will be the consequence; for the labourer, according to this supposition, may be forced by his necessity to consume fewer articles. (H, 239)

Long-run Neutrality of Money Thornton makes clear, however, that the effects just described are of secondary importance. In the long run, the appropriate assumption is the neutrality of money with regard to real economic activity.

> There seems to be only two modes in which we can conceive the additional paper to be disposed of. It may be imagined either, first, to be used in transferring an encreased quantity of articles, which it must, in that case, be assumed that the new paper itself has tended to create; or, secondly, in transferring the same articles at a higher price. Let us examine the first of these cases. . . . When the Bank of England enlarges its paper, it augments, in the same degree, as we must here suppose, its loans to individuals. These favored individuals immediately conceive, and not without reason, that they have obtained an additional though borrowed capital, by which they can push their own particular manufacture. . . . it does not occur to them that the commerce or manufactures of other individuals can be at all reduced in consequence of this encrease of their own. But, first, it is obvious that the antecedently idle persons to whom we may suppose the new capital to give employ are limited in number; and that, therefore, if the encreased issue is indefinite, it will set to work labourers of whom a part will be drawn from other and, perhaps, no less useful occupations. (H, 235-36)

> There remains, therefore, no other mode of accounting for the uses to which the additional supply of it [Bank of England paper] can be turned than that of supposing it to be occupied in carrying on the sales of the same, or nearly the same, quantity of articles as before, at an advanced price, the cost of goods being made to bear the same, or nearly the same, proportion to their former cost, which the total quantity of paper at the one period bears to the total quantity at the other. (H, 241)

Aggregate Demand and the Interest Rate Thornton spends considerable time discussing the economy's aggregate supply function in order to establish both the transitory nonneutrality of money and the long-run neutrality of money with respect to real economic activity. He spends less time discussing the economy's aggregate demand function, apart from a general description of aggregate nominal demand as dependent upon the quantity of money (H, 117-18). There is, however, a section in one of his speeches before Parliament in 1811 in which he explains the relationship between investment demand and the real rate of interest. Thornton first explains the relationship between the real rate of interest and the market rate of interest and then notes that investment demand depends upon the former variable. [This discussion anticipated that of Irving Fisher. See Beranek, Humphrey, and Timberlake (1985).]

> It was material to observe that there had, since the beginning of the war, been a continual fall in the value of money. . . . which was, on the average, 2 or 3 per cent. per annum: it followed from hence that if, for example, a man borrowed of the Bank 1000 pounds in 1800 and paid it back in 1810 . . . he paid back that which had become worth less by 20 or 30 per cent. than it was worth when he first received it. He would have paid an interest of 50 pounds per annum for the use of this money; but if from this interest were deducted the 20 or 30 pounds per annum, which he had gained by the fall in the value of the money, he would find that he had borrowed at 2 or 3 per cent., and not at 5 per cent. as he had appeared to do. . . .

> . . . during a fall in the price of money . . . [men] felt . . . the advantage of being borrowers. . . . on estimating the value of those commodities in which they had invested their borrowed money, they found that value to be continually increasing, so that there was an apparent profit over and above the natural and ordinary profit on mercantile transactions. This apparent profit was nominal, as to persons who traded on their own capital, but not nominal as to those who traded with borrowed money. . . . This extra profit was exactly so much additional advantage . . . and was so much additional temptation to borrow. Accordingly, in countries in which the currency was in a rapid course of depreciation, supposing that there were no usury laws, the current rate of interest was often . . . proportionably augmented. (H, 335-36)

6. The Supply of Money

A Central Bank Thornton developed the conception of a central bank after observing the financial panic of 1793. In particular, he noticed that to banks and to London merchants Bank of England notes were interchangeable with gold specie (H, 123). He argues that in the case of a bank run, the Bank of England should increase its notes in order to offset the reduction in bank reserves caused by gold outflows from the banking system.

> . . . the holder of a note of 1000 pounds . . . carries it to the Bank and demands 1000 pounds in gold. The Bank gives the gold; which gold . . . fills a void in the circulation of the country occasioned by the withdrawing of country bank notes in consequence of alarm, or serves as an

> addition to the fund of country banks. . . . The Bank, therefore, having paid away this 1000 [pounds] in gold, and having received for it their own note for 1000 pounds must now re-issue this note, if they are resolved *to maintain the amount of their paper circulation.* How, then, is the Bank to issue it? The only means which the Bank, on its part, is able to take for the extension of its paper circulation is to enlarge its loans. [Italics in original] (H, 125)

In defending the Bank of England's decision in 1797 to suspend convertibility, Thornton argues that the increase in the Bank's loans that occurred at the time was due to a need to maintain the currency in the face of an internal drain. The increase did not cause the suspension through overissue. "The largeness of those loans was not the *cause* of the guineas going from them, as has been ordinarily supposed; it was the *effect*" [Italics in original] (H, 137).

Thornton discusses monetary base creation by the Bank of England in terms of the Bank's balance sheet (H, 136). He shows its balance sheet as comprising credits of bullion and total loans and debits of capital and deposits plus notes. In Thornton's words, "Every additional loan obtained by the Bank, if we suppose its gold to remain the same, implies an encreased issue of paper" (H, 227).

The Bank Rate—Natural Rate Model of the Money Supply Central to Thornton's theory of money stock determination is the concept of a natural rate of interest, that is, a rate of interest invariant in the long run to the behavior of the money stock. The following quotation asserts this idea as well as the absence of a liquidity effect on the rate of interest in the long run.

> The reader, possibly, may think that an extension of bank loans, by furnishing additional capital, may reduce the profit on the use of it, and may thus lessen the temptation to borrow at five per cent. It has already been remarked in this Chapter that capital by which term *bona fide* property was intended cannot be suddenly and materially encreased by any emission of paper. That the rate of mercantile profit depends on the quantity of this *bona fide* capital and not on the amount of the nominal value which an encreased emission of paper may give to it is a circumstance which it will now be easy to point out.
>
> I admit that a large extension of bank loans may give a temporary check to the eagerness of the general demand for them. It will cause paper to be for a time over abundant, and the price paid for the use of it, to fall.
>
> It seems clear, however, on the principles already stated, that when the augmented quantity of paper shall have been for some time stationary, and shall have produced its full effect in raising the price of goods, the temptation to borrow at five per cent. will be exactly the same as before; for the existing paper will then bear only the same proportion to the existing quantity of goods, when sold at the existing prices, which the former paper bore to the former quantity of goods, when sold at former prices: the power of purchasing will, therefore, be the same; the terms of lending and borrowing must be presumed to be the same. (H, 255-56)

According to Thornton, money creation depends upon the difference between the Bank of England discount rate and the economy's natural rate of interest:

> It may possibly be thought that a liberal extension of loans would soon satisfy all demands and that the true point at which the encrease of the paper of the Bank ought to stop would be discovered by the unwillingness of the merchants to continue borrowing. In order to ascertain how far the desire of obtaining loans at the Bank may be expected at any time to be carried, we must enquire into the subject of the quantum of profit likely to be derived from borrowing there under the existing circumstances. This is to be judged of by considering two points: the amount first of interest to be paid on the sum borrowed; and secondly on the mercantile or other gain to be obtained by the employment of the borrowed capital. . . . We may, therefore, consider this question as turning principally on a comparison of the rate of interest taken at the Bank with the current rate of mercantile profit. (H, 253-54)

In a discussion of an episode of overissue by the Bank of France, Thornton provides a statement of the condition of monetary equilibrium as equality between the Bank rate and the natural rate.[7]

> The French government having occasion in 1805 for some advances on the security of what they call their anticipations . . . proceeded to discount at the Bank as many securities as were sufficient to supply their occasions. . . . The consequence of this transaction was an

[7] In constructing this framework, Thornton needed to work out the interrelationships between the markets for capital, credit, and the money stock and the simultaneous determination of the rate of interest among these markets. Money stock creation permitted a transitory divergence between the market rate on bank loans and the natural rate on real capital. The originality of Thornton's framework can be seen through a comparison to the state of interest rate theory at the time of the publication of *Paper Credit.* Schumpeter (1954, 720) notes the dominance of Adam Smith's view that the market rate of interest was merely a reflection of the rate of return yielded by the capital stock. There was no mechanism for the money stock to influence the market rate.

Thornton's model became the basis for the loanable funds model of interest rate determination used by neoclassical economists. It seems likely that his model was transmitted to the neoclassical economists by J. S. Mill in his *Principles* text. See, for example, the discussion in Mill (1865, 645-47). The version employed by the neoclassical economists, however, lacked the forcefulness of Thornton's model because it dropped the idea of a central bank with the ability to expand the monetary base. At the center of Thornton's model was the concept of a modern central bank, that is, a bank capable of controlling the monetary base through bookkeeping operations. As noted in the introduction, this concept was not again developed in a significant fashion until Keynes.

> augmentation of the paper of the Bank of Paris; a drain of their cash followed; the diligences were found to be carrying off silver into the departments. . . . The circulating medium of the metropolis had now plainly become excessive. . . . the French over-issue arose from an attempt to turn certain securities into cash at a rate of interest lower than that which was the natural one. . . . (H, 337 and 339)

In assessing the usefulness of the quantity theory framework, the central issue is the direction of causation between the money stock and the price level. "The reader possibly may think that, in treating of this subject, I have been mistaking the effect for the cause, an encreased issue of paper being, in his estimation, merely a consequence which follows a rise in the price of goods, and not the circumstance which produces it" (H, 197-98). In Thornton's analytical framework, where the money stock is endogenously determined, the money stock and the price level are simultaneously determined. The analytical usefulness of the quantity theory then becomes an issue of identification. If differences exist in the determinants of the supply and demand functions for nominal money, then the equation of exchange, which provides for a compartmentalization of factors affecting the supply and the demand for money, is a useful device in understanding the determination of the price level.

In Thornton's framework, nominal money supply depends upon the difference between the market rate (the loan rate of the banking system) and the natural rate. Money demand, in contrast, depends upon the level of the market rate. Given the Bank discount rate, which determines the market rate, real shocks produce different movements in money supply and demand.

The example Thornton uses to illustrate this point involves an exogenous decision by foreign investors to repatriate temporarily capital from Britain. As a result of their actions, the demand for public debt falls and the accompanying depreciation of the real exchange rate stimulates exports. At the given Bank discount rate, credit demands at banks increase and the money supply increases (H, 257). Nothing has happened, however, to increase real money demand, which depends upon real variables like the interest rate and real income that are ultimately independent of the money stock. In consequence, the different changes in nominal money supply and nominal money demand produced by a real shock must ultimately be reconciled by a change in the price level.

Control of Country Bank Circulation Thornton completes his model of money stock determination by showing that the note circulation of the country banks rested on the base of the note circulation of the Bank of England.[8] He applies Hume's price-specie-flow mechanism to England considered as two regions (London and the country) with fixed exchange rates (between Bank of England notes and country bank notes). Given a fixed Bank of England note circulation, country banks could not overissue their notes without producing a balance of payments deficit that would drain their reserves in gold and Bank of England notes.[9]

> . . . let it be admitted, for a moment, that a country bank has issued a very extraordinary quantity of notes. We must assume these to be employed by the holders of them in making purchases in the place in which alone the country bank paper passes, namely, in the surrounding district. The effect of such purchases . . . must be a great local rise in the price of articles. But to suppose a great and merely local rise is to suppose that which can never happen or which, at least, cannot long continue to exist; for every purchaser will discover that he can buy commodities elsewhere at a cheaper rate; and he will not fail to procure them in the quarter in which they are cheap, and to transport them to the spot in which they are dear for the sake of the profit on the transaction. . . . he will, therefore, require to have his country bank note turned into a Bank of England note. (H, 208-9)

7. The Responsibility of the Bank of England to Limit Bank Liabilities

In arguing that the Bank of England controlled the note issue of the banking system and that the Bank should recognize an explicit responsibility for this control, Thornton was challenging adherents of the real bills view. This view derives its intuitive appeal from the association of money creation with credit creation in a fractional reserve system. A real bill was an IOU given to a seller of goods by a middleman who purchased the goods for resale at a later date. In order to receive immediate payment, the original seller of the goods would take the IOU, the trade bill, and discount it at a bank, that is, sell it at a discount from the face value that reflected the

[8] When Thornton wrote, the only bank whose notes circulated in London was the Bank of England. Outside of London, the notes of the country banks circulated. The country banks held Bank of England notes and gold as reserves.

[9] The nominal exchange rate between country bank notes and Bank of England notes equals the product of the real exchange rate between the commodities of the country and the London area and the ratio of the price levels between these two areas. With the nominal exchange rate between these two kinds of notes fixed, and given the real exchange rate, the Bank of England determined the price level in the country by setting the price level in the London area through the control of its note circulation. Given the price level in the country, the note issue of country banks was determined.

interest rate. He would receive a bank note, an IOU from the bank promising to pay gold or legal tender on demand. When the middleman resold the goods to the ultimate purchasers, he would pay off the IOU note, and the total quantity of bank notes would return to its original level. Bank notes arising from these transactions were then viewed as self-liquidating. From the real bills perspective, bank notes arising from the discounting of real bills are instruments of credit extension. Their quantity is limited by the real credit demands of the commercial sector. Thornton summarized this view as follows:

> The encrease of Bank of England paper . . . is the effect and not the cause of an advanced price of commodities. To enlarge the Bank of England notes merely in proportion as safe and real bills are offered in return for them is only to exchange one species of paper for another, namely, Bank of England notes for bills, which, though not so current or so safe as Bank notes, are sufficiently worthy of credit. It is, therefore, simply to afford a guarantee to the transactions of the merchant and thus to render that accommodation to commerce which it belongs to the Bank to give. (H, 230-31)

Real bills proponents argued furthermore that, if currency were overissued, it would not remain in circulation, but rather would be used to pay off loans. [See Humphrey (1982) on the real bills principle.]

Thornton uses his Bank-rate natural-rate model of money stock determination to refute the real bills view "that the Bank paper has a natural tendency sufficiently to limit itself" (H, 232). In his model, the central bank must target some nominal variable like the exchange rate in order to ensure equality between the Bank rate and the natural rate and thus to provide for a well-defined money stock and price level. He provides his most succinct criticism of the real bills view in a criticism of John Law. Thornton argues that the real bills assumption that an excess supply of currency will produce a reduction in the quantity of currency through a liquidation of bank loans fails to understand the price level as a monetary phenomenon. An excess supply of currency can create its own demand through a rise in the price level.

> He [Law] forgot that there might be no bounds to the demand for paper; that the increasing quantity would contribute to the rise of commodities; and the rise of commodities require, and seem to justify, a still further increase. (H, 342)

In parts of *Paper Credit*, Thornton argues that because a variety of factors could cause shifts in velocity, there would be no simple relationship between the money stock and the price level. In order that his discussion not be misconstrued, however, he also emphasizes that variability in the public's demand for money in no way reduces the responsibility of the Bank of England to provide for an explicit limitation on the quantity of money in order to preserve a well-defined price level.

> But although there is so great difficulty in estimating the precise influence on the cost of articles, or on the market price of bullion, which each alternation in the quantity of Bank of England notes may produce, there is no reason, on that account, to doubt the general truth of the proposition . . . that the restriction of the paper of the Bank of England is the means both of maintaining its own value, and of maintaining the value, as well as of limiting the quantity, of all the paper in the country. (H, 225)

8. The International Adjustment Mechanism

Assuming the operation of the international gold standard, Thornton extends the price-specie-flow mechanism, which as exposited by Hume had dealt only with exogenous changes in the money stock, to deal with real disturbances and the consequent monetary repercussions. [See Viner (1924) and (1937).] Thornton also presents the first discussion of the operation of floating exchanges rates in the context of the relationship between the internal and external value of the pound.

International Gold Standard Thornton begins his exposition with an explanation of the self-equilibrating character of the balance of payments. The condition of flow equilibrium in the trade sector is derived from the need for stock equilibrium in the market for money and securities.

> It may be laid down as a general truth that the commercial exports and imports of a state . . . naturally proportion themselves . . . and that the balance of trade . . . cannot continue for a very long time to be either highly favorable or highly unfavorable to a country. For that balance must be paid in bullion or else must constitute a debt. To suppose a very great balance to be paid, year after year, in bullion is to assume such a diminution of bullion in one country, and such an accumulation of it in another, as are not easily imagined. . . . To suppose large and successive balances to be formed into a debt is to assume an accumulation of debt which is almost equally incredible. (H, 141-42)

Thornton also derives aggregate balance in the foreign trade sector from the budget constraints of individuals [Perlman (1986)].

> There is in the mass of the people, of all countries, a disposition to adapt their individual expenditure to their income. Importations . . . are limited by the ability of the individuals of that country to pay for them out of their income. . . . And this equality between private expendi-

> tures and private incomes tends ultimately to produce equality between the commercial exports and imports. (H, 142-43)

Under the assumption of fixed exchange rates, Thornton examines the effects of foreign remittances to subsidize continental governments fighting Napoleon and of bad harvests on the British balance of payments. Although not clearly stated, his argument is that the resulting balance of payments deficit will cause the domestic price level to fall and the foreign price level to rise. The deterioration in the real terms of trade, that is, the rise in the price of foreign commodities in terms of domestic commodities, will eliminate the deficit (H, 145).

Thornton criticizes the antibullionist position that "The evil of an unfavorable foreign exchange, and of a consequent high price of gold, arises from an unfavorable balance of trade and from that cause only" (H, 230-31). That is, Thornton criticizes the antibullionist position that exchange rate movements were due solely to the behavior of excess demand and supply in the foreign trade sector. Thornton argues that excess demand in the trade sector and excess supply in the market for the quantity of money are reflections of each other. In the case of real sector shocks, like poor harvests, the direction of causation runs from excess demand in the trade sector to excess supply in the market for money.

> I conceive, therefore, that this excess [of paper], if it arises on the occasion of an unfavorable balance of trade, and at a time when there has been no extraordinary emission of notes, may fairly be considered as an excess created by that unfavourable balance. (H, 151)

In the case of excess issue of the currency, a monetary shock, the direction of causation runs from excess supply in the market for the quantity of money to excess demand in the trade sector.

> "the coming and going of gold" does not . . . "depend wholly on the balance of trade." It depends on the quantity of the circulating medium issued; or it depends, as I will allow, on the balance of trade, if that balance is admitted to depend on the quantity of circulating medium issued. (H, 248)

Fluctuating Exchange Rates Thornton presents the purchasing power parity doctrine according to which fluctuating exchange rates will vary in order to maintain constant the terms of trade when domestic price levels change.

> . . . as goods are rendered dear in Great Britain . . . our exports will be diminished; unless we assume . . . that some compensation in the exchange is given to the foreigner. . . . our imports also will encrease. . . . these two effects . . . will follow provided that we suppose, what is not supposable, namely, that, at the time when the price of goods is greatly raised in Great Britain, the course of [the] exchange suffers no alteration. . . . The fall in the selling price abroad of bills payable here will operate as an advantage to the foreign buyer of our commodities in the computation of the exchangeable value of that circulating medium of his own country with which he discharges the debt in Britain contracted by his purchase. It will thus obviate the dearness of our articles: it will serve as compensation to the foreigner for the loss which he would otherwise sustain by buying in our market. (H, 198-99)

Thornton argues that a floating exchange rate would maintain equality between exports and imports, although account had to be made for desired capital flows (H, 246-47).

Thornton constructed the analytical apparatus of *Paper Credit* in order to deal with the bullionist-antibullionist debate over the cause of the depreciation of the pound on the foreign exchanges. This debate turned on whether the depreciation of the foreign exchange value of the pound following Britain's suspension of the gold standard was a real or a monetary phenomenon. Under convertibility, the Mint had coined one ounce of gold into 3 pounds 17 shillings 10½ pence. In 1801, the pound price of gold rose above this former mint price. Bullionists argued that the excess of the market price over the mint price of bullion was due to the Bank of England's excess issue of its notes. That is, an increase in the money stock had led to a rise in the price level and also to a rise in the price of the specific commodity, gold bullion. Because Napoleonic Europe was on a gold standard, the pound price of gold bullion was the British exchange rate. Thornton believed that the rise in the price of gold bullion reflected a deterioration in Britain's real terms of trade. What is of enduring interest, however, is not Thornton's specific position in this debate, but rather his development of an analytical framework general enough to explain changes in the exchange rate as either a real or a monetary phenomenon.

Thornton's argument in support of his position that the depreciation of the pound was a real phenomenon possesses two parts. The first part is the presumed absence of monetary disturbances. At the time Thornton lived, the idea of an index number had not been invented, and there were no indexes for the price level. In the absence of evidence on the behavior of the price level, Thornton's argument for the absence of monetary disturbances turns on the other variables in the equation of exchange: 1) the money stock, 2) velocity, and 3) real transactions. 1) In support of their position, the bullionists pointed to an increase in Bank note issue between 1797 and

1801. Thornton disputes this position by arguing that the increase in Bank notes from 1795 to 1801 just offset the reduction in the circulation of gold guineas following the Restriction (H, 214). These guineas had been exported when gold coin ceased to circulate. That is, according to Thornton, the monetary base had remained unchanged. 2) Thornton does admit that the velocity of money had increased due to technological change in the payments industry, especially in the form of clearing houses to facilitate check clearing among banks (H, 101 and 222). 3) Thornton then claims, however, that the growth in British trade abroad that accompanied the continental wars had increased real transactions in Britain by enough to offset the effect on prices of the rise in velocity (H, 221-23).

The second part of Thornton's argument is that, while monetary disturbances appeared to be absent, there were obvious real disturbances that could have affected the real exchange rate, especially, the occurrence of two successive poor harvests that had increased British imports of food (H, 225). Finally, Thornton points out that the transitory nature of the shocks affecting the terms of trade would in time allow for resolution of the dispute over the cause of the depreciation of the pound on the foreign exchanges. If the depreciation were real, then the transitory nature of these real shocks would imply the reversal of the pound's depreciation (H, 221).

9. The Bullion Committee Report

Early in 1809, the gold bullion price of the pound fell sharply in Britain. Under convertibility, the Mint had made 123¼ grains of gold interchangeable with one pound sterling. In 1809, only 107 grains of gold were required to buy a pound. On February 1, 1810, Francis Horner, in the House of Commons, moved to form The Select Committee on the High Price of Bullion. Its report, issued on June 8, 1810, accuses the Bank of England of depreciating the value of the pound on the foreign exchanges through overissue of its notes.

The Antibullionist Arguments With Napoleonic Europe on the gold standard, the pound price of gold bullion measured the exchange rate between Britain and Europe. The antibullionists argued that the rise in the price of bullion reflected a deterioration of the balance of payments. They carried their argument further by contending that the real bills policy precluded the possibility that the note issue of the Bank of England could affect the exchange rate. With a real bills policy, it was argued, an excess supply of Bank notes could not arise.

> The Bank Directors . . . professed themselves to be most thoroughly convinced that there can be no possible excess in the issue of Bank of England paper, so long . . . as the discount of mercantile Bills is confined to paper of undoubted solidity, arising out of real commercial transactions, and payable at short and fixed periods. (C, 46)

> Mr. Whitmore, the late Governor of the Bank, expressly states, "The Bank never forces a Note in circulation, and there will not remain a Note in circulation more than the immediate wants of the public. . . . The Bank Notes would revert to us if there was a redundancy in circulation, as no one would pay interest for a Bank Note that he did not want to make use of. (C, 47)

According to the antibullionists, because there could be no excess supply of money, there could be no relationship between the note issue of the Bank and the value of the pound on the foreign exchanges. Mr. Pearse, Governor of the Bank of England, testified:

> In considering this subject with reference to the manner in which Bank notes are issued, resulting from the applications made for discounts to supply the necessary want of Bank notes, by which their issue in amount is so controlled that it can never amount to an excess, I cannot see how the amount of Bank notes issued can operate upon the price of Bullion, or the state of the Exchanges. (C, 33)

The Bullionist Rebuttal The Bullion Committee noted first that exchange rate movements can have a real, as well as a nominal, component (C, 22, 24, and 26). Its members also recognized the reasonableness of the position that the real terms of trade had depreciated. Its members, however, did not recognize the validity of the antibullionist argument that the real bills principle of the Bank of England precluded the emergence of an excess supply of money that could depreciate the value of the pound in the foreign exchange market. The Bullion Committee argued that the Bank Directors did not understand how suspension of the gold standard removed the institutional mechanism for determining the nominal quantity of money. The real bills principle did not provide for an appropriate check on the money stock because it determined the note circulation on the basis of credit demands.

> So long as the paper of the Bank was convertible into specie at the will of the holder, it was enough, both for the safety of the Bank and for the public interest in what regarded its circulating medium, that the Directors attended only to the character and quality of the Bills discounted, as real ones and payable at fixed and short periods. . . . It was hardly to be expected of the Directors of the Bank that they should be fully aware of the consequences that might result from their pursuing, after the suspension of cash payments, the same system which they had found a safe one before. (C, 48-49)

> . . . while the convertibility into specie no longer exists as a check to an over issue of paper, the Bank Directors have not perceived that the removal of that check rendered it possible that such an excess might be issued by the discount of perfectly good bills. . . . That this doctrine is a very fallacious one, Your Committee cannot entertain a doubt. The fallacy upon which it is founded lies in not distinguishing between an advance of capital to Merchants and an additional supply of currency to the general mass of circulating medium. (C, 50)

Finally, the Bullion Committee argued that, given the usury law existing in Great Britain, only explicit rationing of use of the discount window, not the real bills principle, would limit the money stock (C, 57).

In the absence of an index of the price level, the Bullion Committee argued indirectly that the external depreciation of the pound was caused to a significant degree by overissue of Bank notes. In particular, the Committee argued that both the money stock and the velocity of money had increased. Its estimates of the money stock showed a significant increase beginning in 1809 (C, 62ff.). The Committee members used two arguments to show that the velocity of money had increased. First, they argued that velocity depends positively upon the state of confidence in private credit and that this confidence was high. Second, Committee members argued that technological innovation in the payments industry, in particular, the spread of checks and of clearing houses, had increased velocity.

10. The Committee's Conclusions

The Bullion Committee concluded that the depreciation of the pound on the foreign exchanges was primarily due to overissue by the Bank of England. Its members argued that during the suspension of the gold standard the behavior of the foreign exchange rate should serve as a criterion for setting the quantity of money. The Committee also drew the more fundamental conclusion that Parliament should put in place some institutional arrangement for providing a limitation on the nominal quantity of money (C, 45 and 49).

The Bullion Committee rejected a discretionary approach to monetary policy. It argued that discretionary adjustment of the money stock to changes in the public's demand for money was insurmountably difficult.

> The suspension of Cash payments has had the effect of committing into the hands of the Directors of the Bank of England, to be exercised by their sole discretion, the important charge of supplying the Country with that quantity of circulating medium which is exactly proportioned to the wants and occasions of the Public. In the judgment of the Committee, that is a trust, which it is unreasonable to expect that the Directors of the Bank of England should ever be able to discharge. The most detailed knowledge of the actual trade of the Country, combined with the profound science in all the principles of Money and Circulation, would not enable any man or set of men to adjust, and keep always adjusted, the right proportion of circulating medium in a country to the wants of trade. . . . If the natural system of currency and circulation be abandoned, and a discretionary issue of paper money substituted in its stead, it is vain to think that any rules can be advised for the exact exercise of such a discretion. (C, 52-53)

The Bullion Committee recommended the return to the international gold standard. This standard dictates a country's domestic price level. The quantity of gold a country demands at the given price level is then provided through the balance of payments.

> When the currency consists entirely of the precious metals, or of paper convertible at will into the precious metals, the natural process of commerce, by establishing Exchanges among all the different countries of the world, adjusts, in every particular country, the proportion of the circulating medium to its actual occasions, according to that supply of the precious metals which the mines furnish to the general market of the world. (C, 52-53)

The Bullion Committee also defined for the gold standard the different response of a central bank appropriate to an internal drain and an external drain.

> It appears to Your Committee that the experience of the Bank of England in the years 1793 and 1797, contrasted with the facts which have been stated in the present Report, suggests a distinction most important to be kept in view between that demand upon the Bank for Gold for the supply of the domestic channels of circulation, sometimes a very great and sudden one, which is occasioned by a temporary failure of confidence, and that drain upon the Bank for Gold which grows out of an unfavourable state of the Foreign Exchanges. The former, while the Bank maintains its high credit, seems likely to be best relieved by a judicious increase of accommodation to the Country: the latter . . . ought to suggest to the Directors a question whether their issues may not be already too abundant. (C, 60)

11. Summary

In order to address the policy issue of the cause of the depreciation of the pound on the foreign exchanges, Thornton created an enduring analytical framework. The framework is of a general equilibrium nature and demonstrates the relationship between the internal and external value of money under fixed and floating exchange rates. The framework is capable of distinguishing real from monetary phenomena. In particular, Thornton could use it to consider changes in the exchange rate that were both real and monetary in origin. Thornton's framework contains an aggregate supply function that allowed for the transitory nonneutrality of money and for the long-run neutrality of money. In his later speeches, Thornton also sketched out an aggregate demand function dependent upon the real rate of interest. Thornton's model is a natural rate model, that is, real variables are unaffected by the systematic actions of monetary policy. (An exception, of a second-order of magnitude, is made for the distributional effects due to the seigniorage from money creation.)

Thornton created a quantity theory framework that could incorporate paper money, the fiduciary issue of a fractional-reserve banking system. He had a sophisticated theory of the demand for real money that made the velocity of the components of the money stock vary with the difference between the market rate of interest and the own rate on the particular component. The supply of nominal money depends upon the difference between the market rate and the natural rate of interest. Because the interest rate enters in differently in the supply and demand schedules for nominal money, real sector shocks affect these schedules differently. Because the real variables that affect real money demand are only transitorily related to the nominal money stock, the price level must adjust to equilibrate shifts in the supply and demand schedules for nominal money.

Thornton developed the idea of a modern central bank that exercises control over all the liabilities of commercial banks. Monetary base creation by the Bank of England makes possible the transitory divergence between the natural and market rate of interest that leads to money creation. A recurrent theme in Thornton's work is the responsibility of the Bank of England to provide for explicit limitation of the monetary base in order to ensure a well-defined money stock and price level. During times of financial panic and bank runs, this responsibility requires the Bank of England to expand the monetary base in order to maintain the money stock. Thornton summarizes his policy prescriptions in the following passage:

> To limit the total amount of paper issued, and to resort for this purpose, whenever the temptation to borrow is strong, to some effectual principle of restriction; in no case, however, materially to diminish the sum in circulation, but to let it vibrate only within certain limits; to afford a slow and cautious extension of it, as the general trade of the kingdom enlarges itself; to allow of some special, though temporary, encrease in the event of any extraordinary alarm or difficulty, as the best means of preventing a great demand at home for guineas; and to lean to the side of diminution, in the case of gold going abroad, and of the general exchanges continuing long unfavorable; this seems to be the true policy of the directors of an institution circumstanced like that of the Bank of England. To suffer either the solicitations of merchants, or the wishes of government, to determine the measure of the bank issues, is unquestionably to adopt a very false principle of conduct. (H, 259)

12. Concluding Comment

The major theme in Henry Thornton's *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* is the central bank's responsibility for determining the money stock and the price level. The major theme of the *Bullion Report* is that this responsibility should be made explicit and that the mechanism chosen for determining the price level should not be a matter of ongoing discretion.[10]

The ideas of Henry Thornton continue to challenge the monetary policymaker today. Although it is now recognized that the Federal Reserve System bears the responsibility for the behavior of the price level, the procedure for determining it over time is not explicitly enunciated. The basic issue is what kind of anchor the monetary authority should provide for nominal values. Should this anchor remain fast in the sand so that the wind of real sector and monetary shocks moves the ship of nominal economic values around permanent moorings such as price level stability? Alternatively, should this anchor drag across the sand so that the wind of real sector and monetary shocks moves the ship of nominal values randomly away from any given location?

To repeat, two basic approaches to determining the price level over time are possible. One approach would precommit to a long-run path for the price level and consistently provide for some constraint on each period's decision making in order to assure that over time the price level moves around the given long-run path. The other approach, which is the current one, allows the price level to evolve on an ongoing basis through the accumulation of discretionary decisions made each period so that the price level wanders over time without any fixed point of return. Although close to two hundred years old, Thornton's work continues to challenge the modern policymaker to defend the institutional procedures chosen to anchor the nominal values of the economic system.

[10] These two themes are repeated, respectively, in Black (1986) and Black (1987). On the latter issue, see Broaddus and Goodfriend (1984).

References

Beranek, William, Thomas M. Humphrey, and Richard H. Timberlake. "Fisher, Thornton and the Analysis of the Inflation Premium." *Journal of Money, Credit and Banking* 17 (August 1985): 371-77.

Black, Robert P. "A Proposal to Clarify the Fed's Policy Mandate." *The Cato Journal* 5 (Winter 1986): 787-95.

________. "Inside the Black Box." Speech given to the National Association of Business Economists, San Francisco, June 18, 1987.

Broaddus, Alfred, and Marvin Goodfriend. "Base Drift and the Longer Run Growth of M1: Experience from a Decade of Monetary Targeting." Federal Reserve Bank of Richmond, *Economic Review* 70 (November/December 1984): 3-14.

Great Britain. House of Commons. *Report from the Select Committee on the High Price of Bullion*, 1810. In *The Paper Pound of 1797-1821, The Bullion Report*, edited by Edwin Cannan. New York: Augustus M. Kelley, 1969.

Hicks, John R. "Thornton's Paper Credit" (1802). In *Critical Essays in Monetary Theory*, by John R. Hicks, 174-88. Oxford: Clarendon Press, 1967.

Horner, Francis. Review of *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*, by Henry Thornton. *Edinburgh Review* 1 (October 1802): 172-201. In *The Economic Writings of Francis Horner*. Series of Reprints of Scarce Works on Political Economy, no. 13, edited by Frank W. Fetter. London: London School of Economics and Political Science, 1957.

Humphrey, Thomas M. "Adam Smith and the Monetary Approach to the Balance of Payments." Federal Reserve Bank of Richmond *Economic Review* 67 (November/December 1981): 3-10.

________. "The Real Bills Doctrine." Federal Reserve Bank of Richmond *Economic Review* 68 (September/October 1982): 3-13.

________. "Cumulative Process Models from Thornton to Wicksell." Federal Reserve Bank of Richmond, *Economic Review* 72 (May/June 1985): 18-25.

Hutchison, T. W. "Henry Thornton." In *International Encyclopedia of the Social Sciences*, edited by David L. Sills, 14-17. New York: Macmillan & Free Press, 1968.

Keynes, John Maynard. *A Tract on Monetary Reform* (1923). In *The Collected Writings of John Maynard Keynes*, vol. 4. London: The Macmillan Press, 1971.

McCallum, Bennett T. "Some Issues Concerning Interest Rate Pegging, Price Level Determinacy, and the Real Bills Doctrine." *Journal of Monetary Economics* 17 (January 1986): 135-60.

Mill, John Stuart. *Principles of Political Economy* (1865, 6th ed.). London: Longmans, Green, and Company, 1909.

Mints, Lloyd. *A History of Banking Theory*. Chicago: University of Chicago Press, 1945.

Perlman, Morris. "The Bullionist Controversy Revisited." *Journal of Political Economy* 94 (August 1986): 745-62.

Sargent, Thomas J., and Neil Wallace. "The Real-Bills Doctrine versus the Quantity Theory: A Reconsideration." *Journal of Political Economy* 90 (December 1982): 1212-36.

Schumpeter, Joseph A. *History of Economic Analysis*. New York: Oxford University Press, 1954.

Thornton, Henry. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802) and two speeches (1811). Edited with an Introduction by F. A. v. Hayek. New York: Rinehart & Company, Inc., 1939.

Viner, Jacob. *Canada's Balance of International Indebtedness, 1900-1913*. Cambridge: Harvard University Press, 1924.

________. *Studies in the Theory of International Trade*. New York: Harper 1937. Reprinted New York: Augustus M. Kelley, 1965.

CLASSICAL AND NEOCLASSICAL ROOTS OF THE THEORY OF OPTIMUM TARIFFS

Thomas M. Humphrey

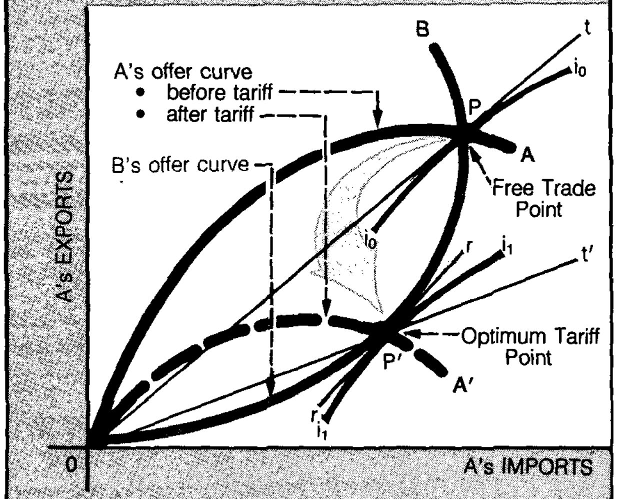
In current debates with protectionists, pure or unilateral free traders insist that unrestricted commerce is optimally advantageous not only for the world as a whole but for any individual nation, even if it practices it alone. From this idea stems the corollary that a country automatically benefits from the unilateral as well as reciprocal elimination of tariffs. If true, it follows that, far from erecting tariffs, a country should immediately dismantle them and enjoy the benefits of international specialization and division of labor even if other nations do not.

In 1940, however, the British economist Nicholas Kaldor challenged these notions by asserting that a tariff always benefits the levying country provided that the duty is not too large, that the country has monopoly power in world markets, and that other countries do not retaliate with tariffs of their own.[1] Kaldor was here advancing the terms-of-trade or optimum tariff argument according to which trade taxes improve the levying country's welfare by turning the commodity terms of trade (relative price at which exports exchange for imports or the quantity of imports bought by a unit of exports) in its favor, thus giving it a better bargain in world markets. By taxing its imports, the country reduces its demand for those goods thus driving down their world price. Similarly, by taxing its exports it lowers the quantity of those goods supplied on the world market thus raising their price. In other words, it acts as a monopolist exploiting an imperfectly elastic foreign supply of its imports or demand for its exports. In so doing it renders its imports cheaper and its exports dearer such that it obtains a larger quantity of imports per unit of exports given up. Of course this terms-of-trade gain comes at the expense of a loss in real trade volume. The optimum rate of the duty is that which maximizes the excess of the gain from terms-of-trade improvement over the loss from lower trade volume and reduced international division of labor.

Kaldor demonstrated these propositions with a geometrical diagram showing the tariff-imposing country choosing to exchange the combination of exports for imports that allows it to reach its highest attainable trade indifference curve given the offer curve of the foreign country (see Figure 1).[2] Shortly after, in 1944, Abba Lerner in his *Economics of Control* described how the same propositions could be illustrated with conventional demand and supply curves (see Figure 2).[3] Both diagrams quickly worked their way into international trade textbooks

Figure 1

Kaldor's Model of the Optimum Tariff



Country A's tariff shifts her offer curve from 0A to 0A' putting her on her highest trade indifference curve i_1 permitted by B's offer curve. We move from point P to P'. Here A's terms of trade improve from 0t to 0t'. The optimum tariff is the wedge between the terms of trade 0t' and the relative value (domestic price ratio) which A's consumers place on the two goods as shown by the slope r of the indifference curve i_1 at point P'.

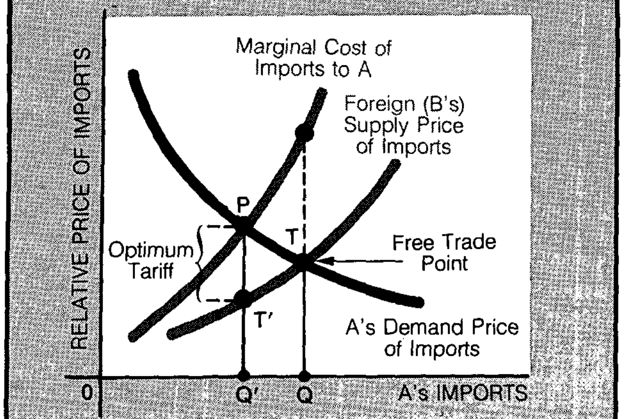
[1] N. Kaldor, "A Note on Tariffs and the Terms of Trade," *Economica*, n.s. 7 (November 1940): 377.

[2] Kaldor, p. 379.

[3] A.P. Lerner, *The Economics of Control* (New York, Macmillan, 1944), pp. 357-59.

Figure 2

Lerner's Demand and Supply Curve Analysis of the Optimum Tariff


RELATIVE PRICE OF IMPORTS
Marginal Cost of Imports to A
Foreign (B's) Supply Price of Imports
P
T
Optimum Tariff
Free Trade Point
T′
A's Demand Price of Imports
0
Q′
Q
A's IMPORTS

Free trade occurs at point T where domestic importers equate import demand price or marginal value with import supply price or the terms of trade. This point, however, is suboptimal for the country because import marginal cost rises faster than supply price and thus exceeds demand price or marginal value. (Marginal cost rises faster than supply price because one must raise price to coax out the last unit supplied and this price rise applies to all previous units. Adding this latter sum to supply price gives marginal cost.) An optimum tariff of PT′ wedged between demand price and supply price lowers the latter thus improving the terms of trade. It also induces domestic importers to limit the quantity of imports to the point where their marginal value to consumers just equals their marginal cost.

where they became the standard model employed in explaining the theory of the optimum tariff. Little was said about earlier work on the subject. From the point of view of the textbooks, the theory to all intents and purposes largely dates from Kaldor's demonstration.[4]

To set the record straight, one must take issue with this view. For, contrary to the impression conveyed by textbooks, optimum tariff theory hardly originated with Kaldor's model but rather long predated it. It can be documented that rudimentary statements of the theory go back at least to the 1830s and 1840s, that these statements were embodied in formal economic models rather than in mere casual remarks, and that virtually all the elements of optimal tariff theory were in place by 1907. In short, the origins of optimum tariff theory are to be found in an earlier vintage of models neglected by the textbooks. A systematic survey of these models helps clarify what economist Murray C. Kemp calls the "confusing and little known early history" of the terms-of-trade argument.[5] It also dispels the notion that all leading classical and neoclassical trade theorists were doctrinaire free traders. True, of the six discussed below, at least four thought that free trade was the best policy from a practical standpoint. On a purely abstract plane, however, all saw the terms-of-trade argument as a valid theoretical qualification to the doctrine that free trade is the best of all possible worlds for each country.

HISTORICAL EVOLUTION

Early optimum tariff models evolved through five distinct stages. First came the demonstration that import duties improve the terms of trade either through gold flows and their effects on relative national price levels or by restricting import demand. Next came the showing that export taxes accomplish the same result by restricting export supply and that the extent of terms-of-trade improvement depends crucially upon the size of certain demand elasticities. There followed a geometrical restatement of these results using the newly developed tool of offer curve analysis. Next appeared indifference curve and consumer surplus models measuring the gain from terms-of-trade improvement and specifying the tariff rate that maximizes the gain. Finally came a mathematical statement of the theory including a rigorous demonstration that a tariff can improve national welfare and a derivation of the formula for the optimum tariff. Each stage saw at least one different innovator—Torrens, Mill, Marshall, Sidgwick, Edgeworth, and Bickerdike being the key names here—advance the theory.

ROBERT TORRENS

Priority for being the first to publish a formal optimum tariff model goes to Robert Torrens in

[4] Texts employing versions of the Kaldor-Lerner model with no mention of its nineteenth and early twentieth century predecessors include R.E. Caves and R.W. Jones, *World Trade and Payments*, 3rd ed. (Boston: Little, Brown, and Co., 1981), pp. 212-13; H.R. Heller, *International Trade: Theory and Empirical Evidence* (Englewood Cliffs: Prentice-Hall, 1968), pp. 145-47; C.F. Kindleberger, *International Economics*, rev. ed. (Homewood, Ill.: R.D. Irwin, 1958), pp. 617-20; D.B. Marsh, *World Trade and Investment* (New York: Harcourt Brace, and Co., 1951), pp. 316-20; and J. Vanek, *International Trade: Theory and Economic Policy* (Homewood, Ill.: R.D. Irwin, 1962), pp. 294-97.

[5] M.C. Kemp, "The Gain from International Trade and Investment: A Neo-Heckscher-Ohlin Approach," *American Economic Review* 56 (September 1966): 788.

1844. Long before then, however, he had perceived that tariffs can turn the terms of trade in favor of the levying country. He stated that idea as early as 1824 in his *Essays on the Production of Wealth* and subsequently elaborated it in a series of letters published in the *Bolton Chronicle* in 1832-33 and reprinted in his 1833 *Letters on Commercial Policy*. Finally, in Letter II and the Postscript to Letter IX of his 1844 *The Budget*, he presented the idea in the form of a hypothetical two-country, two-good model—his famous Cuba case—in which he showed that a 100 percent tariff, via its effect on reciprocal demands, produces an equivalent 100 percent improvement in the terms of trade.[6] This result he depicted in two versions of his model: a monetary version involving specie flows and their effects on local prices and incomes and a pure barter version involving trade in commodities. In the monetary version, terms-of-trade improvement comes from tariff-induced gold movements that raise the price of the protecting country's exports relative to the price of its imports. In the barter version, the same improvement comes from a reduced real demand for imports. Of the two, the monetary version provoked the stronger criticism from Torrens's free trade contemporaries. For that reason, it is described in some detail below.

Torrens's Cuba Model

In the monetary version of his model, Torrens assumed that Cuba specializes in producing sugar and England specializes in cloth, both goods being produced under conditions of constant real costs. He further assumed that each good bears the same duty-exclusive price wherever sold, that the prices of home-produced goods vary directly with the quantity of money in each country, and—of crucial importance to the particular quantitative results he obtained—that each country's demand for the other's export good is of unit elasticity.

Employing these assumptions, he traced a chain of causation from tariff to reduced quantity of imports bought to trade balance surplus to specie inflow and thence to a rise in the price of the protecting country's exports relative to the price of its imports. More precisely, he supposed that, starting from a situation of balanced free trade with England, Cuba imposes a 100 percent ad valorem duty on imports of English cloth. That good being produced at constant cost, the immediate result is to double its price in Cuba causing the quantity demanded to fall by half, the Cuban demand for cloth being assumed by Torrens to be of unit elasticity. In other words, Cubans' total expenditure (price-times-quantity) on taxed cloth remains unchanged; but only half that outlay goes to English exporters, the other half being intercepted by the Cuban government at the customs house.

But these are only proximate or first-round effects. Later-round effects ensue. For, given the volume of Cuban exports, the halving of her import bill produces a favorable trade balance with England and a compensating specie flow from that country lowering general prices in England and raising them in Cuba. Since the price of each country's exportable commodity moves with its general price level and since identical exportable goods bear the same (duty adjusted) price in all markets, the price of sugar rises in Cuba (and England) while the price of cloth falls in England (and Cuba).

The fall in the price of cloth together with the rise in Cuban money incomes occasioned by the specie flow raises the quantity of cloth demanded in Cuba. Conversely, the rise in sugar prices combined with the fall in English incomes reduces the quantity of sugar demanded in England. Gold continues to flow from England to Cuba, lowering incomes in the one and raising them in the other and likewise lowering cloth prices and raising sugar prices, until the resulting stimulus to cloth sales and check to sugar sales restores trade balance equilibrium.

In the new equilibrium, Cuba imports the original quantity of English cloth at two-thirds the original unit price (four-thirds including duty) but exports only half the initial quantity of sugar at four-thirds the initial unit price. In barter terms, Cuba purchases the same real quantity of imports at the cost of only half the initial quantity of exports given up, her commodity terms of trade having improved 100 percent. England's terms of trade of course deteriorate by the same amount.

Barter Version of Torrens's Model

Torrens derived exactly the same results in the pure barter version of his model, which he elaborated with great precision in his Postscript to Letter IX of *The Budget*. There he argued (1) that the equilibrium terms of trade must lie between the comparative cost ratios in the two countries, (2) that the precise location of that equilibrium depends upon each country's reciprocal demand for the product of the other,

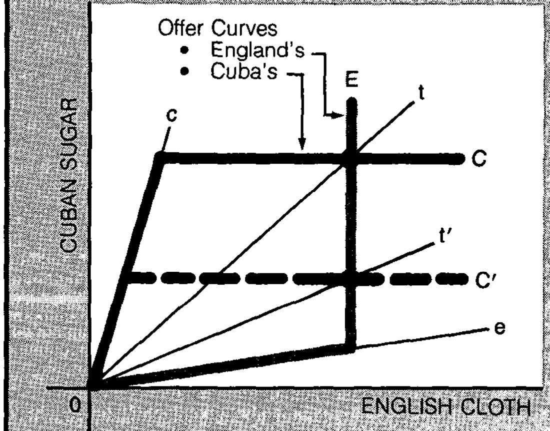
[6] On Torrens's Cuba model, see D.P. O'Brien's *The Classical Economists* (London: Oxford University Press, 1975), pp. 191-94; L.C. Robbins's *Robert Torrens and the Evolution of Classical Economics* (London: Macmillan, 1958), pp. 199-203; and J. Viner's *Studies in the Theory of International Trade* (New York: Harper, 1937), pp. 298-99, 322, 463.

(3) that the resulting equilibrium lies most in favor of the country with the weakest reciprocal demand, and (4) that a tariff, by reducing the levying country's reciprocal demand, turns the terms of trade in its favor. Although he drew no diagrams himself, the essentials of his analysis can be depicted with the aid of Marshallian reciprocal demand or offer curves showing the determination of the equilibrium terms of trade by the intersection of the two curves (see Figure 3).

As drawn, the curves differ from offer curves found in standard textbooks in two respects. First, they bend toward equilibrium only at the points on the respective internal comparative cost ratio lines at which the countries would operate in the absence of trade. Second, within the range at which trade occurs they take the form of horizontal and vertical straight lines reflecting Torrens's assumption of unit elastic reciprocal demands. Given these elasticities and starting from free trade equilibrium, Cuba's tariff shifts her effective offer curve down to half its initial level thus producing at the original terms of trade an excess world demand for sugar and a corresponding excess supply of cloth. To eliminate these excess supplies and demands England's terms of trade deteriorate by 100 percent. In the new equilibrium England imports half the initial quantity of sugar at the cost of the same initial amount of cloth given up. Here is the key idea of optimum tariff models; namely that trade taxes influence reciprocal demands which determine the terms of trade thus allowing governments to manipulate those terms.

Figure 3

Torrens's Tariff Model



The lines 0c and 0e represent the domestic comparative cloth-to-sugar cost ratios in Cuba and England, respectively. They show the high opportunity cost of producing cloth in Cuba and sugar in England in the absence of trade. The assumed constancy of costs makes the lines straight. The lines set the outer limits for the offer curves 0C and 0E because neither country will trade at terms worse than it can obtain under autarky.

The offer curves follow the comparative cost lines over a range in which the countries are indifferent to trade. Then the offer curves depart from those lines, indicating each country's willingness to offer exports in exchange for imports at different terms of trade.

The horizontal and vertical segments of the offer curves reflect Torrens's assumption of unit elastic reciprocal demands. In other words, the curves depict the case in which the quantity of imports demanded by each country varies equiproportionally with changes in the terms of trade such that the country always offers a constant quantity of exports in exchange.

Cuba imposes a 100 percent tariff on English cloth. Cuba's offer curve shifts from 0C to 0C'. Her terms of trade improve from 0t to 0t'.

Money Stock Implications

The foregoing terms-of-trade effects were important. To Torrens, however, they were overshadowed by the impact of Cuba's tariff on England's money stock. In the monetary version of his model he explains how the redistribution of specie occasioned by the tariff produces a one-third expansion of Cuba's money stock and a corresponding one-third contraction of England's. No country, he thought, could endure a monetary contraction of such magnitude. For the resulting collapse of product prices would bring ruinous rises in the real burden of debts, wages, taxes, and other fixed charges whose nominal values are sticky and thus respond sluggishly to deflationary pressure. Economic stagnation, "national bankruptcy, and revolution would be the probable results."[7]

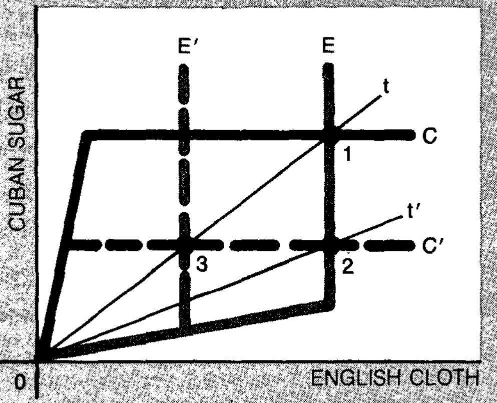
Reciprocity in Commercial Policy

Having shown how England might lose from foreign tariffs, Torrens next used his analysis to argue for reciprocity in tariff removal. He pointed out (1) that a unilateral abolition of tariffs would, like their foreign imposition, worsen the home country's terms of trade and reduce its money stock, (2) that equal retaliatory duties would cancel the unfavorable terms-of-trade and monetary effects of foreign levies, and (3) that the simultaneous removal of duties by all countries tends to leave money stocks and the terms of trade unchanged (see Figure 4). On these grounds

[7] R. Torrens, Letter II of *The Budget. On Commercial and Colonial Policy.* (London: Smith, Elder and Co., No. 65, Cornhill, 1844), p. 37.

Figure 4

Torrens on Trade Policy Reciprocity



Cuba's tariff lowers her offer curve and improves her terms of trade. Trade equilibrium goes from point 1 to point 2. England's countervailing duty shifts her offer curve leftward and corrects her terms-of-trade deterioration (point 2 to point 3).

The same process works in reverse. Unilateral tariff removal by England shifts her offer curve rightward and worsens her terms of trade (point 3 to point 2). Reciprocal tariff removal by Cuba, however, restores the terms of trade to its original level (point 2 to point 1). Simultaneous tariff removal by both countries achieves the same result instantaneously (point 3 to point 1).

Moral: Reciprocity leaves the terms of trade unchanged.

he proposed that Britain counter foreign tariffs with equal duties of her own, that she trade freely only with countries admitting her goods duty free, and that she drop her tariffs only insofar as her trading partners abolish theirs.

Criticisms

Torrens's analysis was unsympathetically received by his contemporaries who feared it would undermine the case for free trade. His critics refused to accept policy conclusions drawn from a two-by-two model regarded by them as an inaccurate representation of a world economy characterized by many goods and many countries. Herman Merivale argued that competition from third countries producing sugar for export would limit Cuba's power to manipulate the terms of trade.[8] Also England could avoid Cuba's tariff by selling to third countries and exporting goods other than taxed cloth, such alternatives being possible in a multi-good, multi-country model. This point was made by Nassau Senior who also noted that what Cuba gains through terms-of-trade improvement might be outweighed by her loss of productivity and competitiveness due to reduced international specialization and division of labor.[9] The most cogent criticism, however, came from George Warde Norman. He noted that England's terms of trade would hardly deteriorate to the extent claimed by Torrens if one dropped the assumption of unit elastic demands. He also argued that the logic of Torrens's model implied that England should levy not equal but higher tariffs than those levied abroad to improve the terms of trade and that such action would intensify the danger of a trade war with all parties losing.[10] These criticisms were telling. For Torrens indeed had overlooked the possibility of trade warfare and the likelihood that highly elastic reciprocal demand schedules would in the long run severely limit the effectiveness of tariffs.

JOHN STUART MILL

Although Torrens's Cuba case was the first optimum tariff model to appear in print, it was hardly the first formulated. Already in 1829-30, some fifteen years earlier, John Stuart Mill had constructed a similar model which he subsequently presented in the first of his *Essays on Some Unsettled Questions in Political Economy*, a volume he published in 1844 in response to Torrens's *The Budget.*

Mill's model possessed most of the features of the monetary version of Torrens's Cuba model, namely two countries, two goods, complete specialization, constant costs, law of one price, Hume's price-specie-flow mechanism, and quantity theory of money. But Mill greatly enriched the model by permitting demand elasticities to range from zero to infinity and by incorporating export as well as import taxes into the analysis. In so doing, he expanded the model's explanatory power thus enabling it to cover a greater variety of cases than con-

[8] H. Merivale, *Lectures on Colonization and Colonies*, II, 1842, pp. 308ff. On Merivale's criticisms see Viner, *Studies*, p. 322 and Robbins, *Robert Torrens*, pp. 209-11.

[9] N. Senior, "Free Trade and Retaliation," *Edinburgh Review* 88 (July 1843): 12-15, 29-35. On Senior's analysis see O'Brien, *The Classical Economists*, pp. 194-95.

[10] G.W. Norman, *Remarks on the Incidence of Import Duties with special Reference to the England and Cuba Case contained in "The Budget,"* privately printed (London: T. and W. Boone, 29 New Bond Street, 1860), pp. 8, 12-19. On Norman's criticisms, see O'Brien, *The Classical Economists*, pp. 195-96.

sidered by Torrens. In particular, he showed how different elasticities affect the degree of terms-of-trade improvement.

Export Taxes, Foreign Demand Elasticities, and the Terms of Trade

Mill applied his model first to export taxes, concluding that such taxes tend to improve the taxing country's terms of trade by an amount equal to, more than, or less than the tax as the elasticity of the foreign demand for exports is equal to, less than, or greater than one.[11] To demonstrate, he employed an example in which England exports cloth to and imports linen from Germany. In his example, he assumed that England levies a tax on her exports of cloth to Germany. Cloth being produced in England at constant real cost, its price to Germans rises initially by the amount of the tax. Provided the German demand for cloth is of unit elasticity such that her import expenditure on that good remains unchanged after the tax raises its price, there results no disturbance to the balance of payments requiring equilibrating specie flows and further adjustments in the prices of the traded goods. Cloth prices paid to England consequently remain above their pre-duty levels by exactly the amount of the tax. And there being no change in the price of England's import good (linen), her terms of trade—that is, the ratio of the price of cloth to the price of linen—improves exactly by the amount of the tax. In short, unit elastic German demand ensures a terms-of-trade improvement equiproportional to the tax.

On the other hand, if Germany's demand for English cloth is inelastic such that she spends more on that good when the tax boosts its price, her import bill will rise producing a deficit in her trade balance. The resulting flow of specie from Germany to England will, via the operation of the quantity theory of money and the law of one price, raise further the price of cloth and lower the price of linen in both countries. England will purchase more of the cheaper linen and sell less of the dearer cloth, these demand readjustments acting to restore trade balance equilibrium. In the new equilibrium, England receives a price for her cloth raised by more than the tax. As she will also be paying a lower price for German linen, her terms of trade—the relative price of cloth exports to linen imports—will have improved by more than the tax.

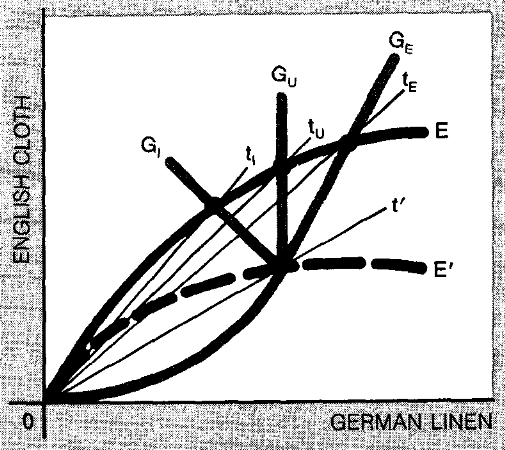
Finally, if Germany's demand for English cloth is elastic such that she spends less on it when the tax raises its price, her import bill will shrink producing a surplus in her trade balance and a corresponding specie flow from England. The result of this money flow is to lower the world price of cloth and to raise the world price of linen—these price changes continuing until cloth sales are stimulated and linen sales checked sufficiently to restore trade balance equilibrium. With England's export prices somewhat lower than they were immediately after the imposition of the tax and her import prices somewhat higher, her terms of trade have improved but by less than the amount of the tax.

Mill's Model in Barter Terms

The foregoing conclusions can be presented in barter terms, although why Mill himself did not do so is something of a mystery since he applied barter analysis involving his notion of reciprocal demand schedules to other problems of trade theory. In any case, Figure 5 shows England's terms of trade improving in greater, equal, or lesser proportion to the export tax as the German offer curve is backward bending (i.e., inelastic), vertical (of unit elasticity), or upward sloping (elastic), respectively—just as Mill's monetary model predicts.

Figure 5

Mill's View of the Size of Terms-of-Trade Improvement and the Elasticity of the Foreign Offer Curve



England's export tax shifts her offer curve from OE to OE'. The resulting terms-of-trade improvement is proportionately larger than, equal to, or smaller than the tariff as the German offer curve is inelastic (backward bending), unit elastic (vertical), or elastic (upward sloping).

[11] J.S. Mill, *Essays on Some Unsettled Questions in Political Economy* (1844), (London: London School of Economics and Political Science, 1948), pp. 21-24.

An Exception

Mill admitted but one exception to the rule that export taxes improve the taxing country's terms of trade: the case of an elastic German demand for cloth combined with an inelastic English demand for linen. Here the specie flow from England caused by the tax-induced decline in Germany's spending on cloth is not self-correcting but rather is self-reinforcing. For the faster gold flows abroad to raise the price of German linen, the more England spends on that commodity. And the more she spends, the greater her loss of gold and the greater the resulting fall in the price of her cloth. To restore equilibrium, cloth may have to fall so low in price relative to linen that the terms of trade turn against England by more than the amount of the tax. Such would be the case, Mill thought, should Germany's expenditure on cloth be so insensitive to changes in income that prices alone had to bear the full burden of adjustment.

Import Tariffs and the Terms of Trade

Having examined the terms-of-trade effects of England's export taxes, Mill next turned his attention to her import tariffs.[12] He concluded that they invariably improve her terms of trade except in the singular case of a totally inelastic English demand for German linen. But as long as England's demand is of greater than zero elasticity, quantity of imports demanded falls as the tariff raises price. Since German exporters producing under conditions of constant cost receive a sum equal to the lower (post-tariff) quantity times the old (pre-tariff) price, it follows that England's import bill falls. The resulting gold flow from Germany to England lowers linen's supply price and raises the price of cloth, thus improving England's terms of trade. No such improvement would occur, however, if England's import demand were perfectly inelastic such that the quantity of linen demanded by that country remained unchanged when the tariff raised its price. With no shrinkage in quantity demanded, the price-times-quantity sum paid to German exporters would be the same as before, which means that there would be no disturbance to the balance of payments requiring gold flows and hence no changes in the absolute and relative prices of cloth and linen. In other words, England's import bill, and hence her terms of trade, would remain unchanged in this case.

Views on Tariff Policy

To summarize, Mill, like Torrens, had clearly established the theoretical possibility of a country improving its terms of trade and its welfare through trade restriction. Interestingly enough, however, Mill opposed the application of his optimum tariff theory to commercial policy on practical and moral grounds. Tariffs, he said, invite retaliatory duties that not only nullify the initial terms-of-trade improvement, but also bring costly reductions in the volume of world trade.[13] Even in the absence of retaliation, tariffs are unjust because one country's gain is another's loss. Moreover, as the rest of the world's loss exceeds the dutying country's gain the tariff is inimical to global welfare and cannot be justified from a cosmopolitan point of view. In his words, "if international morality . . . were rightly understood and acted upon, such taxes, as being contrary to the universal weal, would not exist."[14] He did, however, agree with Torrens that reciprocity was a prime consideration in the decision to remove tariffs. "A country," he said, "cannot be expected to renounce the power of taxing foreigners, unless foreigners will in return practice towards itself the same forbearance. The only mode in which a country can save itself from being a loser by the revenue duties imposed by other countries on its commodities, is to impose corresponding revenue duties on theirs."[15]

ALFRED MARSHALL AND HENRY SIDGWICK

In the 1870s and 1880s Alfred Marshall and Henry Sidgwick constructed optimum tariff models. Marshall's innovation was to transform Mill's model into geometry, expressing his results in terms of reciprocal demand or offer curves showing each nation's desired quantity of exports and imports as a function of the terms of trade. Sidgwick too expressed some of Mill's conclusions in purely barter terms, but without adding much to his analysis.

Marshall, in an unpublished manuscript which Professor John Whitaker dates at 1872-74, employed his reciprocal demand curves to show that when both curves are elastic (provided the foreign curve is not infinitely so) a tax on imports or exports always im-

[12] Mill, pp. 26-27.

[13] Mill, pp. 28-29.

[14] Mill, p. 25.

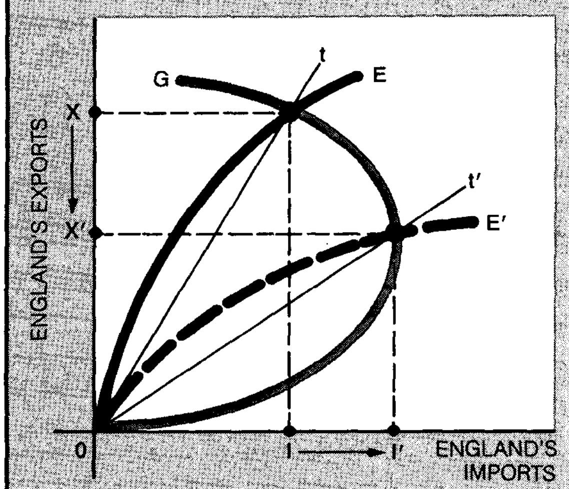
[15] Mill, p. 29.

proves the terms of trade of the levying country.[16] He also showed that when the foreign curve is inelastic—meaning that the foreign country offers a greater total quantity of its exports as its terms of trade deteriorate—then the dutying country enjoys a two-fold gain.[17] Not only do its terms of trade improve, but, by obtaining a larger total quantity of imports and sacrificing a smaller total quantity of its exports, it has more of both goods to consume at home (see Figure 6). A country lucky enough to face an inelastic foreign offer curve, said Marshall, has nothing to lose and everything to gain by exploiting it.

In general, however, Marshall thought that the ability of the taxed country in a multi-country, multi-commodity world to switch its production to non-taxed exports and to trade its goods in nontaxed markets rendered its offer curve so highly elastic as to leave the dutying country little scope for tariff-induced improvements in the terms of trade. He also feared that the pressure of special interests would push tariff rates far above the optimum level such that the dutying country as well as the whole world would lose.

Figure 6

Marshall on the Gains from Exploiting an Inelastic Foreign Offer Curve



England imposes either an import tariff or an export tax that shifts her offer curve from 0E to 0E'. Because Germany's offer curve is inelastic or backward bending, England enjoys a two-fold gain. Her terms of trade improve from 0t to 0t'. Also she sacrifices fewer of her exports while obtaining a larger quantity of imports. Thus she has more of both goods to consume at home.

Sidgwick's analysis closely followed that of Marshall, from whose abandoned 1873-77 manuscript on trade theory Sidgwick had printed for private circulation selected chapters under the title *The Pure Theory of Foreign Trade* (1879). In particular, Sidgwick stressed three points previously made by Marshall. First is the importance of monopsony power in achieving terms-of-trade improvement. No country, he said, could expect to improve its terms of trade by means of tariff unless it "supplied a considerable part of the whole demand for the [taxed] foreign products."[18] Second, a tariff affects the terms of trade through its impact on reciprocal demands. Specifically, A's tariff reduces her demand for B's good, thus producing an excess world supply of that good. This excess supply is only eliminated by a deterioration in B's terms of trade.

> Supposing trade to be in equilibrium at the time that the demand in A for B's commodities is artificially restricted by import duties raising their price, and supposing that other things—*including the demand in* B for A's commodities—remain unchanged, one obvious result will be that B will import more than she exports; hence in order to restore the balance of trade, a certain readjustment of prices will be necessary by which B will in most cases tend to obtain a somewhat smaller aggregate of imports on somewhat less advantageous terms.[19]

Third, the effectiveness of A's tariff depends upon the elasticity of B's offer curve. If that curve is almost totally inelastic, as when B urgently requires A's good at any price, the terms-of-trade gain realized by A comes at the cost of little or no shrinkage in her export volume. But if B's offer curve is perfectly elastic, as when she can readily substitute third-country goods for A's good in her consumption mix, A's tariff will have no effect other than diminishing her (A's) real trade volume. Said Sidgwick:

> This restriction on B's import trade may possibly not reduce materially the amount of her imports from A, if the commodities supplied by A are strongly demanded in B . . . . On the other hand . . . if the products of A are closely pressed in the markets of B by the competition of other countries, the protection given by A to . . . her industry may very likely have the secondary effect of inflicting a blow upon . . . the exports from A to B.[20]

Here is Sidgwick's recognition of one point stressed by optimum tariff theory, namely that a tariff is powerless to improve the terms of trade when the foreign offer curve is perfectly elastic.

[16] J.K. Whitaker (ed.), *The Early Economic Writings of Alfred Marshall, 1867-1890*, Vol. 1 (New York: Free Press, 1975), p. 270.

[17] Whitaker, pp. 275-76.

[18] H. Sidgwick, *The Principles of Political Economy*, 2nd edition (London: Macmillan and Company, Ltd., 1887), p. 492.

[19] Sidgwick, pp. 494-95.

[20] Sidgwick, p. 495.

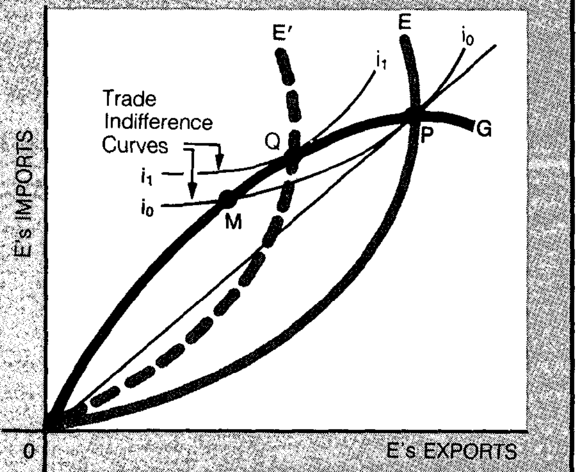
FRANCIS Y. EDGEWORTH

Although Torrens, Mill, Marshall, and Sidgwick had shown that tariffs could benefit the dutying country by turning the terms of trade in its favor they did not provide a measure of this benefit nor did they specify the precise tariff rate that would maximize it. Not until 1894 did these ideas make their first appearance with the publication of F. Y. Edgeworth's famous *Economic Journal* article on "The Pure Theory of International Values." There in a demonstration that anticipated Kaldor's in all essential respects, he employed the now-standard curves of trade geometry to identify the optimum tariff (see Figure 7). In so doing he advanced the theory in at least four ways.

First, he superimposed on Marshall's reciprocal demand or offer curves trade indifference curves essential to the demonstration of welfare gains from trade restriction. His diagram shows the home country's trade indifference curve i_0 passing through the free trade point P at which the offer curves intersect the (free-trade) terms-of-trade line.[21] This particular indifference curve, he said, indicates the level of welfare or satisfaction the home country enjoys under free trade. It provides a benchmark against which to compare alternative welfare levels yielded by different degrees of trade restriction.

Figure 7

Edgeworth's Optimum Tariff Model



England's free trade indifference curve is i_0. A movement to any position on the foreign (German) offer curve between points P and M would put England on a higher indifference curve representing an improvement in her welfare. England wants to reach the highest indifference curve permitted by the foreign offer curve. The optimum tariff is that which shifts England's offer curve to 0E' such that point Q is attained.

Second, he specified the range of tariff rates beneficial to the home country. To do so, he noted that the same indifference curve that passes through the free-trade point P also cuts the foreign offer curve at point M, which, by virtue of being on the same indifference curve, yields the same level of welfare as the free trade point. Since all points on the foreign offer curve between these two extremes lie on higher indifference curves, it follows that any movement to a position between points P and M will result in the home country being better off than under free trade. In other words, points P and M mark the range of terms-of-trade improvement beneficial to the home country. Somewhere within this range benefit is at a maximum.

Third, he identified the point Q at which the home country reaches its highest possible trade indifference curve given the foreign offer curve. The optimum tariff, said Edgeworth, is that which distorts the home country's offer curve such that it intersects the foreign offer curve at this point of tangency with the highest attainable indifference curve. Here, almost fifty years before Kaldor himself presented it, is the famous tangency solution to the determination of the optimum tariff.

Fourth, Edgeworth showed that if the tariff is raised too much it reduces rather than increases welfare. For as the tariff is raised from point P to Q to M, welfare at first rises, reaches a maximum, and starts to fall. And if the tariff is raised beyond point M, welfare falls below the level attained at the free trade position P. It follows that the tariff must not be too large if the nation is to benefit.

Finally, he noted some pitfalls in the practical application of the model. For one thing, the optimum point, though precisely identifiable in theory, cannot be ascertained with any accuracy in practice. Another consideration is the strong political pressure exerted by protectionists. These factors make it all too likely that policymakers would raise tariffs far beyond the optimum point thus lowering welfare. Then too there was the likelihood of retaliation which would nullify any gains generated by the tariff. Above all was the immorality of tariffs from the cosmopolitan point of view; there is little to be said for restrictions that cause other countries to lose more than the dutying country gains.[22] Taking all

[21] F.Y. Edgeworth, "The Theory of International Values, II," *Economic Journal* 4 (September 1894): 432. The same diagram appears in his *Papers Relating to Political Economy*, Vol. 2, (London: Macmillan, 1925), p. 39.

[22] Edgeworth, *Papers* II, pp. 17 (n.5), 18.

these factors into account, free trade, he thought, remains hands down the best and most practical policy for a nation to follow.

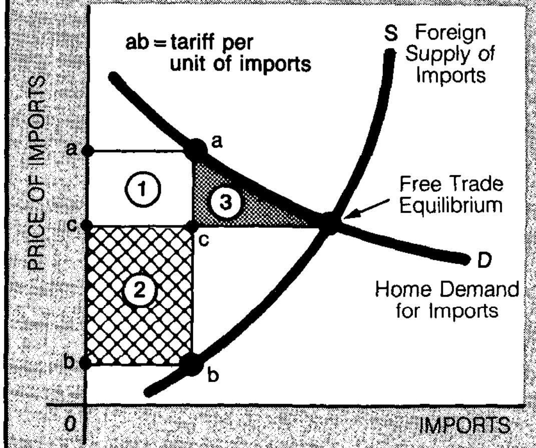
C. F. BICKERDIKE

The last economist to be considered is C. F. Bickerdike, who in his 1906 *Economic Journal* article on "The Theory of Incipient Taxes" and his 1907 Review of A. C. Pigou's *Protective and Preferential Import Duties* contributed at least four innovations to optimum tariff theory. First, he emphasized the similarity between the theory of monopoly and the theory of tariffs. He noted that when an individual exporter expands his sales he drives down the price received by other exporters. An export tax, he claimed, corrects this tendency for competition among exporters to lower the price obtained by all. It does so by extracting from the gross price received by exporters the amount by which an extra unit sold lowers the price on all previous units. In so doing, the duty forces exporters to behave as if they take account of their collective influence on prices paid by foreigners. The result is that the country acts as a single monopoly unit that fully exploits its bargaining power to improve the terms of trade.[23] In effect, the export tax acts to form competing exporters into a cartel.

Second, he specified anew the welfare gain from trade restriction. As an alternative to Edgeworth's indifference curve measure, he defined the net benefit of an import duty as the sum of the tax revenue collected from foreigners through lower import prices less the deadweight loss in consumers' surplus caused by the shrinkage in trade volume. This welfare gain he illustrated in a Marshallian demand-and-supply curve diagram (see Figure 8) in which crosshatched rectangular area 2 measures tariff revenue collected from foreigners and shaded triangular area 3 is the deadweight loss in consumers' surplus.[24] To avoid Torrens-Mill type specie flow and price level movements—complications that could shift the demand and supply curves in Figure 8—he assumed that each country operated with an inconvertible paper currency of constant purchasing power. As noted by John Chipman, this assumption effectively transformed a partial equilibrium diagram into a consistent general equilibrium model.[25]

Figure 8

Bickerdike's Measure of Net Gain from Tariff



A tariff of amount ab wedged between import demand and supply lowers the quantity imported, raises import demand price from c to a, and reduces import supply price from c to b. Government collects tariff revenue equal to the sum of rectangular areas 1 and 2. Of these two areas, the first represents tariff revenue paid by domestic consumers in the form of higher prices. The second, however, represents revenue paid by foreign suppliers in the form of lower prices received. As the portion of the tariff paid by domestic consumers is either returned to them in the form of lower taxes or spent for their benefit, the loss in consumers' surplus is limited to triangular area 3. The net gain from the tariff is equal to the crosshatched rectangle 2 minus the shaded triangle 3. As long as the former exceeds the latter in size the tariff improves national welfare.

In any case, Bickerdike concluded from his diagram that a tariff benefits the dutying country whenever rectangle 2 exceeds triangle 3 in size, which will be the case provided the tariff is small enough, the demand curve is of greater-than-zero elasticity, and the import supply curve is not infinitely elastic. He also concluded that the tariff is more beneficial the more elastic the levying country's demand for imports. This is true because the more elastic the demand curve the larger the foreigner's tariff-burden rectangle relative to the deadweight loss triangle. In the limiting case of infinitely elastic demand, foreigners would bear the entire burden of the tariff and deadweight loss would be zero.

Third, he provided the first mathematical proof that a country could gain from a tariff. To obtain his proof he constructed a two-country, two-commodity algebraic model consisting of five groups of equations. These included (1) export and import demand and

[23] C.F. Bickerdike, "The Theory of Incipient Taxes," *Economic Journal* 16 (December 1906): 530-31.

[24] Bickerdike, p. 533-34.

[25] J.S. Chipman, "Bickerdike's Theory of Incipient and Optimal Tariffs," unpublished paper, 1987.

supply functions, (2) a trade-balance equilibrium condition, (3) a law-of-one-price equation stating that the foreign exchange rate must be such as to equalize the common currency price (tariff-adjusted) of each good across countries, (4) a tariff equation defining the percentage tariff wedge inserted between the prices domestic importers pay and foreign suppliers receive, and (5) a collective utility function defining national welfare as the excess of the total utility from consuming import goods over the cost of obtaining that utility through the production of exports.[26] Having constructed his model, he then had to demonstrate that national welfare increases upon a small increase in the tariff. This he accomplished by substituting equations (1) through (4) into the utility function, differentiating that function with respect to the tariff, and then showing that the resulting first derivative is positive. His expression reveals the welfare gain as depending critically upon export supply and import demand elasticities at home and abroad.

Last but not least he expressed the optimum tariff rate in terms of a mathematical formula, being the first to do so. To derive his optimum tariff formula he set the foregoing first derivative of utility with respect to the tariff rate equal to zero as required for a maximum and solved for the tariff rate (or more precisely for the reciprocal of one plus the tariff rate—this term being his measure of the tariff wedge). The result was

$$T = \frac{1-(1/\eta_\delta)}{1-(1/\eta_\sigma)}$$

where T is the reciprocal of one plus the optimum tariff rate t, or $1/(1+t)$, and η_δ and η_σ denote the export demand and import supply elasticities of the foreign country.[27] Solving this formula for the tariff rate t yields the expression

$$t = \frac{1/\eta + 1/\epsilon}{1 - 1/\epsilon}$$

where $\eta = -\eta_\sigma$ and $\epsilon = \eta_\delta$. Here is the classic formula for the optimum tariff later made famous by R. F. Kahn, J. de V. Graaff, and Harry G. Johnson in the 1940s and 1950s.[28]

CONCLUSION

The impression conveyed by textbooks notwithstanding, economists hardly had to wait until the 1940s to obtain theoretical models of the optimum tariff. On the contrary, the key components of such models already had been assembled long before. Robert Torrens in the 1840s supplied two elements, namely the notions that reciprocal demands determine the terms of trade and that tariffs affect those reciprocal demands thus giving policymakers a means of manipulating the terms of trade. John Stuart Mill showed in an essay published in 1844 that an export tax works as well as an import tariff to improve the terms of trade and that the extent of the improvement depends crucially on the size of the coefficients of elasticity of demand. Alfred Marshall in the 1870s translated the Torrens-Mill analysis into graphical form thus establishing the reciprocal demand or offer curves used in modern models of the optimum tariff. To Marshall's reciprocal demand schedules Edgeworth in 1894 added trade indifference curves thus allowing one to identify in principle the particular tariff rate that maximizes national gain. Finally, C. F. Bickerdike in the early 1900s added three more components to the theory: he proved mathematically that a tariff could improve national welfare, he presented alternative measures of the resulting gain, and he derived the algebraic formula for the optimum tariff rate. He also showed that the optimum tariff restrains competition among individual importers and exporters so that the dutying country acts as a cartel exploiting its market power to improve the terms of trade.

Except for Torrens and Bickerdike, these same economists also specified the basic shortcomings of optimum tariff theory. The theory, they noted, assumes unrealistically (1) that foreign countries will not retaliate with tariffs of their own, (2) that elasticities of supply and demand in foreign trade are not so large in the long run as to render the tariff ineffective, (3) that the optimum tariff rate can be precisely identified and skillfully administered, and (4) that politicians can resist pressures to raise tariff rates above the optimum level. None of these assumed conditions, they felt, were likely to be realized in practice. They further pointed out that a tariff can benefit no nation except at the cost of greater injury to others and is thus unacceptable from a cosmopolitan point of view. For these reasons they remained convinced that, despite the theoretical

[26] Bickerdike, Review of *Protective and Preferential Import Duties* by A.C. Pigou, *Economic Journal* 17 (March 1907): 100.

[27] Bickerdike, Review, p. 101.

[28] See R.F. Kahn, "Tariffs and the Terms of Trade," *Review of Economic Studies* 15, no. 1 (1947): 16; J. de V. Graaff, "On Optimum Tariff Structures," *Review of Economic Studies* 17, no. 1 (1949): 53; and H. G. Johnson, "Alternative Optimum Tariff Formulae," in his *International Trade and Economic Growth* (London: George Allen and Unwin, 1958), p. 60.

case that could be made for an optimum tariff, free trade was the best policy from a practical and moral standpoint. In other words, they established virtually all the arguments for and against the use of an optimum tariff long before modern analysts rediscovered the issue. Here is a prime example of classical and neoclassical economists formulating theories relevant to current trade policy analysis.

NEW PUBLICATIONS

MONETARY POLICY IN PRACTICE

Marvin Goodfriend

The Federal Reserve Bank of Richmond is pleased to announce the publication of *Monetary Policy in Practice*. The author combines analytical, quantitative, and historical methods with institutional knowledge of the Federal Reserve to study monetary policymaking as it relates to financial markets, bank regulation, and international finance. Written for university professors, undergraduate and graduate students, and financial market participants. Copies can be obtained free of charge by writing to Public Services Department, Federal Reserve Bank of Richmond, P. O. Box 27622, Richmond, Virginia 23261.

BUYING TREASURY SECURITIES AT FEDERAL RESERVE BANKS
Twelfth Edition

James F. Tucker

This easy-to-read booklet outlines the step-by-step procedure whereby individuals can purchase Treasury securities from the Federal Reserve Banks. In addition, the booklet describes the various types of securities—bills, notes, and bonds—available for purchase. Suitable for the public. $2.00 per copy. Advance payment is required by check or money order in U. S. dollars, payable to the Federal Reserve Bank of Richmond. Send your order and payment to Public Services Department, Federal Reserve Bank of Richmond, P. O. Box 27622, Richmond, VA 23261.

A REVIEW OF BANK PERFORMANCE IN THE FIFTH DISTRICT, 1986

John R. Walter and David L. Mengle

Despite the heightened competition and continuing adjustments resulting from the first full year of regional interstate banking, Fifth Federal Reserve District[1] commercial banks produced during 1986 the highest return on assets and return on equity of the past ten years. On average, Fifth District banks' earnings reached $1 for every $100 in assets and $15.87 for every $100 of equity. The increased profitability was a stark contrast to the decline to 63 cents for each $100 of assets and $10.22 for each $100 of equity experienced by the average of all banks in the United States.

A closer look at the numbers, however, reveals that the rise in return on assets was due to gains from the sale of securities, gains that were the result of falling interest rates. When securities gains are excluded, Fifth District return on assets actually declined from 1985. Since banks nationwide also benefited from securities gains, excluding such gains makes their decline in return on assets even greater. Thus, while gains on securities sales help explain why Fifth District profitability improved, they do little to explain the continuing difference between Fifth District profitability and that of the average of all U.S. banks.

As interest rates fell during 1986, District banks' net interest margin declined significantly. The decline was largely offset by lower provision for loan and lease losses and noninterest expenses. U.S. banks on average also had a decrease in net interest margin, but had higher provision for loan and lease losses and noninterest expenses as well.

Fifth District banks allowed their capital ratios to decline during 1986 so that at the end of 1986 capital adequacy was diminished somewhat. In contrast, U.S. banks reported increased capital ratios on average.

All ratios and figures in this article are based on book values of liabilities and assets. When interest rates vary or the economic health of borrowers declines, book values do not automatically reflect the changes. While book values are probably the best measures available to the public at this time, it is important to be aware of their limitations. The limitations are discussed in the box beginning on page 32 of this article.

Profits

Return on Assets Net income grew by almost 19 percent at Fifth District banks from 1985 to 1986. Table I shows that return on assets (ROA) increased from .98 percent in 1985 to 1.00 percent in 1986.[2] For all U.S banks, net income fell by 1.2 percent, leading to a decline in ROA from .70 percent in 1985 to .63 percent in 1986 (see Appendix). About 8 percent of Fifth District banks reported losses in 1986, while almost 20 percent of all U.S. banks reported losses for the same period. While the average return on assets and return on equity (ROE) for Fifth District banks reached historically high levels for the year, the average figures for all U.S. banks fell to their lowest levels during the period recorded in the Appendix table.

The improvement in gains on securities relative to average assets from 1985 to 1986 figured importantly in the improved profitability of Fifth District banks. Gains or losses on securities are realized when banks sell securities at prices different from their book values. Since interest rates fell during most of 1986, selling securities produced gains. Because securities gains and losses are considered to arise from factors largely outside the control of management, however, they are often excluded from ROA. Excluding them

William Whelpley, formerly of the Federal Reserve Bank of Richmond, provided yeoman's service in the construction of the data base for this article. Frank Fry, Board of Governors of the Federal Reserve System, supplied helpful advice for retrieving the data.

[1] Maryland, the District of Columbia, Virginia, North Carolina, South Carolina, and most of West Virginia. At the end of 1986, there were 605 commercial banks in the Fifth District. During 1986, 25 new banks were established while 21 were merged into other banks for a net gain of four from 1985. No Fifth District commercial banks failed in 1986.

[2] See the definition and discussion of return on assets and return on equity in the box on page 32.

Table I

INCOME AND EXPENSE AS A PERCENT OF AVERAGE ASSETS[1] FIFTH DISTRICT COMMERCIAL BANKS, 1979-86

Item	1979	1980	1981	1982	1983	1984	1985	1986
Gross interest revenue	8.49	9.46	11.15	10.86	9.58	10.02	9.48	8.51
Gross interest expense	4.53	5.60	7.29	6.93	5.82	6.33	5.70	4.97
Net interest margin	3.96	3.86	3.86	3.93	3.76	3.69	3.78	3.54
Noninterest income	0.80	0.90	1.01	1.03	1.16	1.15	1.22	1.22
Loan and lease loss provision	0.26	0.26	0.25	0.28	0.25	0.33	0.46	0.40
Securities gains[2]						-0.02	0.06	0.15
Noninterest expense	3.24	3.37	3.48	3.53	3.45	3.37	3.40	3.29
Income before tax	1.26	1.13	1.14	1.15	1.22	1.12	1.20	1.23
Taxes	0.28	0.20	0.19	0.18	0.22	0.19	0.22	0.23
Other[3]	-0.04	-0.04	-0.09	-0.10	-0.02	0.00	0.00	0.00
Return on assets[4]	0.94	0.89	0.86	0.87	0.98	0.93	0.98	1.00
Cash dividends declared	0.30	0.32	0.33	0.37	0.34	0.31	0.31	0.34
Net retained earnings	0.64	0.57	0.53	0.50	0.64	0.62	0.67	0.66
Return on equity[5]	13.51	12.79	12.56	13.12	15.21	14.62	15.41	15.87
Average assets ($ millions)	80,671	88,280	97,217	108,439	121,173	137,131	156,574	181,133

Note: Discrepancies due to rounding error.

[1] Average assets are based on fully consolidated volumes outstanding at the beginning and at the end of the year.

[2] Banks were required to report securities gains or losses above the tax line on their income statements for the first time in 1984.

[3] Includes securities and extraordinary gains or losses after taxes, for 1979-83 data, and extraordinary items and other adjustments after taxes for 1984-86 data.

[4] Return on assets is net income divided by average assets.

[5] Return on equity is net income divided by average equity. Average equity is based on fully consolidated volumes outstanding at the beginning and at the end of the year.

Source: Consolidated Reports of Condition and Income.

changes the picture for the Fifth District (Table II). Rather than increasing, ROA net of securities gains and losses fell in the Fifth District from .92 in 1985 to .85 in 1986. For all U.S. banks, the measure fell from .64 in 1985 to .50 in 1986. Note that while excluding securities gains affects the comparison between performance in 1985 and in 1986, it does not affect the comparison between Fifth District banks and their peers nationwide.

Chart 1 shows ROAs (including securities gains) for three size classes of Fifth District banks. Only large banks (more than $750 million in 1986 total assets) improved their ROAs in 1986. Large banks' ROAs averaged .97 percent in 1986 compared with .92 percent in 1985. ROA declined to 1.10 percent for medium Fifth District banks (1986 total assets between $100 million and $750 million) and to 1.17 percent at small banks (less than $100 million in total assets). Table II shows that excluding securities gains from ROA leads to decreases for all the size classes. Large banks' ROA less securities gains fell from .85 in 1985 to .79 in 1986, while that for medium banks fell from 1.13 to 1.03 and that for small banks from 1.19 to 1.09.

There were several other factors influencing the changes in ROA for the three size classes in the Fifth District. Net interest margins declined as a percent of average assets for all three. For large banks, lower loan and lease loss provisions added to ROA, while lower noninterest expenses did so for medium-sized banks. For small banks, higher loan and lease loss provisions and lower noninterest revenue helped move ROA down from the preceding year.

Comparing Chart 2 with Chart 1 reveals a striking difference between the performance of small banks nationwide and those in the Fifth District. While Fifth District small bank ROA has remained high throughout the years shown, U.S. small banks' average ROA began falling in 1981 and has dropped each year since then. The decline in profitability outside the Fifth District is largely due to smaller banks' exposure to geographically limited problems such as those in agriculture and the oil industry.

Table II

PROFITABILITY MEASURES BEFORE AND AFTER ADJUSTMENT FOR SECURITIES GAINS AND LOSSES FIFTH DISTRICT COMMERCIAL BANKS

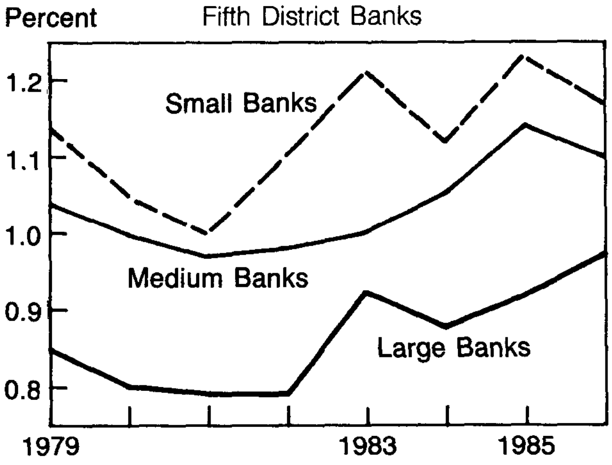
	Small	Medium	Large	Total
1986				
ROA	1.17	1.10	0.97	1.00
ROE	12.59	14.01	16.99	15.87
Sec. gains/losses	0.08	0.07	0.18	0.15
Adjusted ROA	1.09	1.03	0.79	0.85
Adjusted ROE	11.68	13.06	13.88	13.44
Book value leverage	10.80	12.69	17.57	15.80
1985				
ROA	1.23	1.14	0.92	0.98
ROE	13.53	14.99	15.95	15.41
Sec. gains/losses	0.04	0.01	0.07	0.06
Adjusted ROA	1.19	1.13	0.85	0.92
Adjusted ROE	13.08	14.87	14.74	14.50
Book value leverage	10.98	13.11	17.42	15.68

Note: Adjusted ROA (ROE) is net income less securities gains and losses divided by average assets (equity). Leverage is average assets divided by average equity. Discrepancies are due to rounding error.

Looked at differently, in 1981 about 5 percent of all small U.S. banks had negative ROAs, while by 1986 more than 21 percent had moved into the loss column. In the Fifth District, in contrast, about 6 percent of small banks had ROAs below zero in 1981, but this had risen to just 10 percent by 1986. At the other end of the profitability spectrum, in 1981 approximately 60 percent of small U.S. banks had ROAs greater than 1.0 percent. But by 1986 only about 35 percent of small banks could make this claim. In the Fifth District, less than 50 percent of small banks had ROAs over 1 percent in 1981, but this rose to 54 percent by 1986.

Chart 1

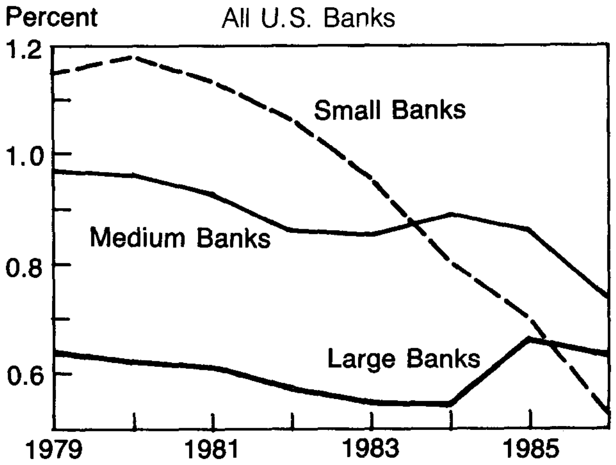
RETURN ON ASSETS*



* Net income divided by average assets.

Chart 2

RETURN ON ASSETS*



* Net income divided by average assets.

Return on Equity Fifth District banks as a group increased average ROE from 15.41 percent in 1985 to 15.87 percent in 1986 (Table II). The increase reflected both higher ROA and increased leverage[3] at large banks. Chart 3 shows, however, that the average ROE performance conceals the performance of small- and medium-sized banks. Excluding securities gains as in Table II would lead to a decline in ROE for all three size classes and for the average of the District.

On average, U.S. banks experienced a decline in ROE from 11.33 percent to 10.22 percent during 1986. Since leverage at the national level was virtually unchanged from 1985, the lower ROE simply reflects the lower ROA for all U.S. banks.

During 1986 Fifth District banks lowered their ratio of retained earnings to average assets from .67 percent to .66 percent. Along with the higher ROA, this enabled banks to increase dividends from .31 percent to .34 percent relative to average assets (Table I). Banks at the national level maintained cash dividends at .33 percent of average assets and allowed retained earnings to decline to .31 percent.

[3] For a definition and discussion of leverage, see page 32.

PROBLEMS IN MEASURING BANK PERFORMANCE

ROA or ROE? The two standard measures of bank profitability are return on assets (ROA) and return on equity (ROE).* ROA is defined as net income as a percent of average assets held during a year, that is,

$$\text{ROA} = \frac{\text{Net Income}}{\text{Average Assets}} \ .$$

Average assets is the average book value of assets held at the beginning and at the end of the year. It is used because income is earned throughout the year and assets are likely to vary during the year. ROE is defined as net income as a percent of average book value of equity outstanding during a year, or

$$\text{ROE} = \frac{\text{Net Income}}{\text{Average Equity}} \ .$$

Which is the more useful profitability measure? As will be seen, neither sums up everything.

ROA is the more straightforward measure of profitability because it shows profits as a yield on all a bank's sources of value. ROA is often used to compare the performance of one bank with another or with the average of all banks. Despite its popularity, there are pitfalls in using ROA to compare banks in a particular year.

For example, a high ROA could be the result of efficient operations or a low cost deposit base. Importantly, it could also be the result of lending at high rates to risky borrowers. It is even possible that a small bank with low overhead expenses and access to low cost deposits might earn a high ROA by placing a high percentage of its funds in securities rather than loans. While gross returns may be lower on securities than on loans, so also may be losses and administrative expenses.

A low ROA could stem from heavy reliance on purchased funds or relatively high cost time deposits. It could also come from conservative lending policies that yield relatively low rates in the current year but fewer loan and lease loss provisions in future years.

ROA reflects the return to both depositors and owners. In contrast, ROE tells how effectively a bank's assets are being used to produce income for its owners. As a guide to the profitability of a bank as an investment, then, ROE appears to be more useful than ROA.

Unfortunately, ROE also has pitfalls when used to compare banks. The definitions of ROA and ROE given above imply the following relationship:

$$\text{ROE} = \text{ROA} \times \text{Average Assets/Average Equity}$$

where the ratio of assets to equity measures the *leverage* of a bank. The relationship implies two things. First, ROE will be subject to the same disadvantages as ROA. Second, differences in leverage between banks will affect their relative ROEs.

Leverage measures how much of a bank's assets is financed by persons other than the equity owners. Because higher leverage means relatively more fixed claims on a bank's assets by depositors, higher leverage means higher *financial risk* to the owners. In other words, leverage magnifies the effect on ROE of changes in ROA. For example, if two banks have the same positive ROA, the more leveraged will have the higher ROE. Similarly, losses will cause ROE to fall more for highly leveraged banks.

Further, large banks tend to be more highly leveraged than small banks. It does not follow, however, that large banks are necessarily more risky than small banks. While higher leverage implies higher financial risk, it says nothing about *business risk* arising from banks' loan, investment, and funding decisions. In fact, a large bank's financial risk from leverage could be more than offset by lower business risk from a more diversified loan and investment portfolio. Unfortunately, neither ROA nor ROE can on its own disentangle the risk components.

There are fewer problems with using ROA and ROE to compare bank performance over time. A group of banks is subject to common influences over time, for example, changes in interest rates and in the fortunes of regional economies. Changes in ROA and ROE would express how a bank or group of banks responded to the common influences. In addition, since banks following risky loan policies might also have higher loan losses over time, differences in risk between banks are more likely to cancel out. That is, a bank following a high risk strategy may report higher net income (and ROA) now, but may have to set aside higher provision for loan and lease losses (and report lower net income and ROA) in later periods.

Still, there are difficulties in making comparisons over time. For example, ROA and ROE could be driven up or down by gains or losses from the sale of securities. In this case, ROA and ROE changes are more the result of interest movements and timing of securities sales than of credit or operational factors under the control of management. For this reason, some analysts exclude securities gains and losses when calculating ROA and ROE (see Text, Table II).

Book Value Accounting A significant problem with ROA, ROE, and other performance measures is that they are calculated from book values of assets, liabilities, and equity. Book values fail to account for changes in the value of assets, liabilities, and equity occurring between their placement on the books of the

* Higgins (1984, chap. 2) discusses in more detail the inherent difficulties of various profitability measures.

bank and their removal by sale, repayment, maturity, or charge-off. The failure of book values to reflect such changes in worth is a serious problem when interest rates fluctuate and when the ability of borrowers to repay debts comes into question.

The data used in the preparation of this article and in most investigations of bank performance come from the Consolidated Reports of Condition and Income (known as call reports) collected from banks by their principal regulators. The reports include a balance sheet, an income statement, and some supporting documents, all of which must be furnished quarterly by every insured bank. Call reports require banks to report book values of assets and liabilities, although market values of securities are also reported.

The book value of a loan, for example, is the amount of money originally advanced or paid for the loan minus principal repayment and minus losses that have been charged off. In the case of securities and loans purchased by the bank at other than par value, book value is the price paid for the security or loan plus an amount to account for the amortization of premium or the accretion of discount. Book value, then, is the historic value of an asset or liability.

If banks used market value accounting, they would value their assets and liabilities at the price, or best estimate of the price, they would trade for in the market. Estimates of market value can be very accurate if there is a developed market for an asset. If there is no active or developed market for a particular asset, however, estimates of market value can be difficult. Still, even rough estimates of market value would probably provide more useful information than book values.**

The biases inherent in book value accounting may show up in different ways. For example, suppose a bank buys a security with a yield of 10.5 percent. If market interest rates fall to, say, 9.5 percent, the market price of the security rises. Under market value accounting, reported net income in the current period would rise by the amount of the price increase, but the return on the security in the current and subsequent periods would decrease to the market level of 9.5 percent. In contrast, under book value accounting the value of the security is not adjusted, so the income from the security continues at the now above-market rate of 10.5 percent. Unless the bank sells the security and recognizes the capital gains, reported income in subsequent periods will be biased upwards from true economic income. Similarly, if interest rates rise the bank avoids having to book the loss, but subsequent reported returns from the security will be biased downward until the security is sold.

A more serious example of disadvantages of book value accounting arises with problem loans. Under current practices a problem loan may be carried at book value so long as it is expected to eventually be paid back in full. In practice, a banker might not set aside reserves on a problem loan unless pressured to do so by regulators. This leads to curious effects on reported ROA and ROE. By failing to adjust the reported value of a loan for anticipated losses, the bank manager avoids having to reduce current reported net income by the amount of the provision for loan and lease losses. Since the loan is a problem loan, however, current income from the loan is probably below its contracted amount. The result is an ROA biased downward from market levels. So, in this case the price of avoiding reduced current income from setting aside loss reserves is lower ROA in subsequent periods. Once a loan is written down to its estimated market value, return on the loan goes back to market levels.

A final problem with using book values in measuring performance arises because the value of equity outstanding is reported at book value rather than market value. Thus, ROE does not express profitability as yield realized in the market by investors. Rather, book ROEs may be biased upward or downward from actual market yields depending on whether the shares of the bank would sell below or above their book values. In addition, leverage measures based on book values may give a distorted picture of a bank's true capital structure.

There are two problems with using market values of equity to compute ROE and leverage, however. First, because the shares of most banks are not actively traded, there are few market transactions from which values could be inferred. Second, most actively traded shares are those of bank holding companies rather than banks, so the market value of the equity might reflect the value of several subsidiary banks as well as nonbank subsidiaries. Thus, even if one wished to use market value of equity to compute ROE and leverage, the required information might not be readily available for all banks.

** For a discussion of the feasibility of market value accounting as a substitute for book value accounting, see Lereah (1986). For arguments in favor of market value accounting, see Benston et al. (1986, chap. 8).

References

Benston, George J., Robert A. Eisenbeis, Paul M. Horvitz, Edward J. Kane, and George G. Kaufman. *Perspectives on Safe and Sound Banking: Past, Present, and Future.* Cambridge, Massachusetts: MIT Press, 1986.

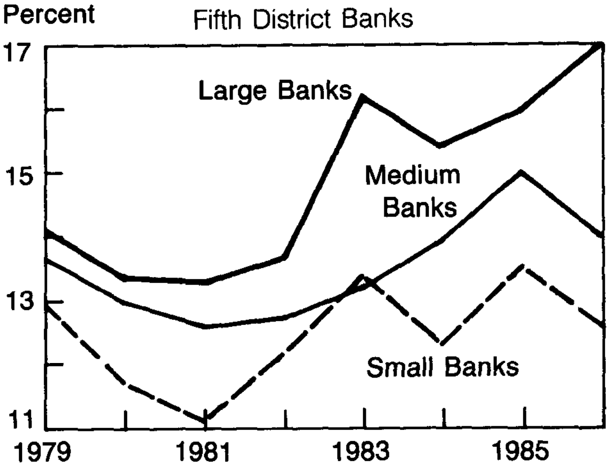
Higgins, Robert C. *Analysis for Financial Management.* Homewood, Illinois: Irwin, 1984.

Lereah, David. "Current-Value Accounting: Feasibility for Financial Institutions." In *Proceedings of a Conference on Bank Structure and Competition*, pp. 311-22. Federal Reserve Bank of Chicago, 1986.

Chart 3

RETURN ON EQUITY*

Fifth District Banks



* Net income divided by average equity.

Interest Margin

Net interest margin, the difference between interest income and interest expense as a percent of average assets, fell 6 percent at Fifth District banks during 1986 (Table I). The decline brought net interest margin to a historically low level. At all U.S. banks, net interest margin fell by 4 percent. Even with the greater decline, Fifth District banks enjoyed a 13 percent higher net interest margin than their U.S. counterparts. Net interest margin declined for all size classes of Fifth District banks, but the decline was greatest at medium-sized banks.

Falling interest rates during 1986 pushed down earnings on assets. Table III shows that earnings from loans and leases declined more than earnings from securities. The ratio of short-term maturity loans and leases to all Fifth District loans and leases was greater than the ratio of short-term securities to all securities, so the loan and lease portfolio of Fifth District banks was more rate-sensitive than the securities portfolio. The decline in market rates therefore led to a greater decline in return on loans and leases than in return on securities.

Fifth District banks' gross interest expense ratio (interest expense as a percentage of average assets) declined by 73 basis points from 1985 to 1986. This decline, like that in interest income, was largely caused by falling interest rates. Table IV shows lower costs of all major categories of liabilities.

Noninterest Revenue and Expense

On average, Fifth District banks experienced no change in noninterest income as a percent of average assets from 1985 to 1986. Service charge and leasing income fell, while other noninterest income grew by an offsetting amount.[4] At the same time, District banks were able to decrease noninterest expenses relative to average assets by 11 basis points during the year. The major part of this decrease came from a decline in salaries expense. While employment by Fifth District banks actually increased by 3 percent in 1986, the number of employees per million dollars of assets fell by 11 percent.

The average results for Fifth District banks conceal differences between size classes. Small banks' decline in service charges and other noninterest income was more than offset by lower salaries and other noninterest expense. Medium banks' fall in service charge income was swamped by a decrease in salaries and bank premises expense. Large banks had stable noninterest income categories but a decline in salaries expense.

[4] Other noninterest income includes such items as income from credit card fees, fiduciary activities, mortgage loan servicing fees, and safe deposit box rentals. Other noninterest expense includes such items as insurance premiums, legal fees, advertising, and charges resulting from litigation or other claims.

Table III

AVERAGE RATES OF RETURN ON SELECTED INTEREST-EARNING ASSETS FIFTH DISTRICT COMMERCIAL BANKS, 1979-86

Item	1979	1980	1981	1982	1983	1984[2]	1985[2]	1986[2]
Total interest-earning assets	10.09	11.28	13.18	12.68	11.11	11.77	11.06	9.78
Total loans and leases	11.25	12.50	14.48	14.14	12.38	12.59	11.92	10.63
Net loans and leases[1]	11.37	12.63	14.64	14.30	12.53	12.74	12.08	10.77
Total securities	6.43	7.15	8.57	9.27	9.20	9.68	9.01	8.30

[1] Net loans and leases are: total loans net of allowance for loan losses for 1979-83; total loans and leases net of the sum of allowance for loan and lease losses and allocated transfer risk reserve for 1984-86.

[2] Total and net loans and leases here include leases while in other columns they do not.

Table IV

AVERAGE COST OF FUNDS FOR SELECTED LIABILITIES
FIFTH DISTRICT COMMERCIAL BANKS, 1979-86

Item	1979	1980	1981	1982	1983	1984	1985	1986
Interest-bearing deposit accounts	7.15	8.68	10.63	9.91	8.19	8.72	7.89	6.77
Large certificates of deposit	9.96	11.33	14.35	12.05	7.62	9.47	7.91	7.07
Deposits in foreign offices	10.28	13.17	15.18	12.79	7.73	9.19	7.92	6.40
Other deposits	6.16	7.54	9.23	9.12	8.34	8.55	7.97	6.74
Subordinated notes and debentures	8.19	8.20	8.11	8.34	8.32	8.03	9.64	8.48
Fed funds	11.94	13.34	15.54	11.21	8.52	9.58	7.67	6.92
Other	6.98	8.65	13.49	11.29	8.75	9.18	6.73	5.19
Total	7.60	9.13	11.23	10.10	8.24	8.84	7.90	6.76

Loan and Lease Loss Provision[5]

Fifth District banks set aside income equivalent to .40 percent of average assets as provision for loan and lease losses, a decrease from the .46 percent set aside in 1985. As in previous years, 1986 Fifth District provision was well below that of banks nationwide. On average, all U.S. banks set aside income equivalent to .76 percent of average assets for provision in 1986 compared with .66 percent in 1985.

Chart 4 shows that in the Fifth District changes in loan and lease loss provision varied considerably with size of bank. Large banks were able to lower their provision during 1986 to .42 percent of average assets. Medium banks' provision was unchanged between 1985 and 1986 at .30 percent of average assets, while small banks increased their provision to .39 percent of average assets.

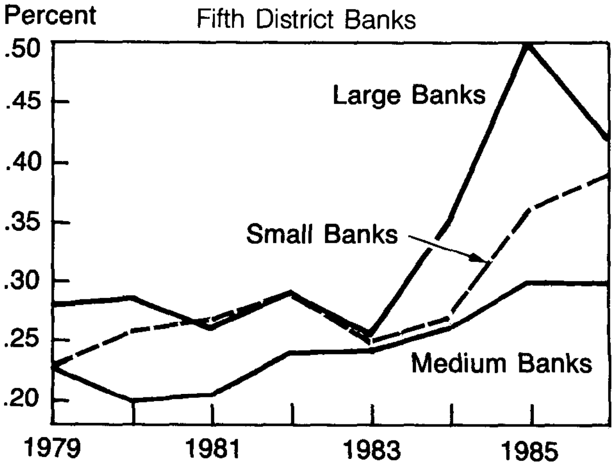
The ratio of nonperforming loans and leases[6] to total loans and leases, and the ratio of loans and leases charged off (net of recoveries) to the total of loans and leases, are measures of the quality of past credit decisions. Historically Fifth District banks have had much lower levels of these ratios than the average for all U.S. banks.

In the Fifth District, nonperforming loans and leases were 1.1 percent of total loans and leases at the end of 1986, unchanged from the previous year. Net charge-offs increased over the period from .41 percent of total loans and leases to .47 percent. District banks apparently set aside sufficient provision to keep allowance at about the same level relative to total loans and leases in 1986 as it had been in 1985.

For all banks in the nation, 2.8 percent of loans and leases were nonperforming, up from 2.7 percent in 1985. Charge-offs rose from .81 percent of loans and leases in 1985 to .93 percent in 1986. In addition, banks at the national level increased allowance as a percentage of total loans and leases.

Chart 4

LOAN AND LEASE LOSS PROVISIONS AS A PERCENT OF AVERAGE ASSETS



[5] Loan and lease loss provision is the income statement flow that adds to the balance sheet stock known as allowance for loan and lease losses. Provision for allocated transfer risk is included in provision for loan and lease losses, and allocated transfer risk reserve is included in allowance for loan and lease losses (except when computing capital ratios).

[6] A nonperforming loan or lease is defined in this article as one that has not been charged off but is 90 days or more past due or is not accruing interest. Net charge-offs are loan and lease losses, net of loans and leases recovered, actually charged against the allowance for loan and lease losses. In other words, they are flows subtracted from the allowance.

Capital

In 1986 the average Fifth District bank's capital ratio fell slightly from its 1985 level (Table V). Looking at each size class shows, however, that only large banks' capital ratio fell while small- and medium-sized banks increased capital rapidly. Specifically, large banks' primary capital to assets ratio declined from 7.04 percent at year-end 1985 to 6.91 percent at the end of 1986. Even with the decline the capital ratio of large Fifth District banks is well above what it was at the end of 1984.

At the national level banks increased their capital ratios on average. Increases took place in large- and medium-sized banks' ratios, while for small banks capital ratios declined slightly. Because U.S. banks have been increasing capital ratios on average for the last three years, by 1986 the national average surpassed that for Fifth District banks. Still, small- and medium-sized Fifth District banks maintained significantly higher capital ratios on average than their peers at the national level.

The components of large Fifth District banks' primary capital that fell (relative to assets) were common stock, capital surplus, and allowance for loan and lease losses. These declines were offset to some extent by increases in undivided profits and mandatory convertible debt. Small and medium Fifth District banks improved their capital ratios by adding to common stock, surplus, undivided profits, and allowances for loan and lease losses. For all U.S. banks, common stock declined relative to assets while surplus, undivided profits, allowance for loan and lease losses, and perpetual preferred stock increased.

Table V

CAPITAL RATIOS
FIFTH DISTRICT AND ALL U.S. COMMERCIAL BANKS

1986

	Small	Medium	Large	Total
Fifth District				
Primary ratio	10.23	8.75	6.91	7.49
Total ratio	10.27	8.77	7.24	7.75
All U.S. banks				
Primary ratio	9.26	8.01	7.03	7.52
Total ratio	9.30	8.15	7.51	7.88

1985

	Small	Medium	Large	Total
Fifth District				
Primary ratio	9.91	8.35	7.04	7.56
Total ratio	9.96	8.40	7.34	7.79
All U.S. banks				
Primary ratio	9.31	7.92	6.84	7.41
Total ratio	9.37	8.10	7.26	7.73

1984

	Small	Medium	Large	Total
Fifth District				
Primary ratio	9.60	8.35	6.64	7.28
Total ratio	9.63	8.41	6.92	7.49
All U.S. banks				
Primary ratio	9.24	7.94	6.35	7.11
Total ratio	9.31	8.15	6.66	7.36

Note: Primary capital here is common stock, perpetual preferred stock, surplus, undivided profits, capital reserves, mandatory convertible instruments,allowance for loan and lease losses, and minority interest in consolidated subsidaries, less intangible assets. Total capital includes primary capital plus limited life preferred stock and those subordinated notes and debentures not eligible for primary capital. Primary capital and total capital are divided by quarterly average assets plus allowance for loan and lease losses less intangible assets to produce primary ratio and total ratio. The measures used here correspond closely but not exactly to the different measures used by the federal bank regulatory agencies.

APPENDIX

INCOME AND EXPENSE AS A PERCENT OF AVERAGE ASSETS ALL U.S. COMMERCIAL BANKS, 1979-86[1]

Item	1979	1980	1981	1982	1983	1984	1985	1986
Gross interest revenue	8.62	9.87	11.81	11.19	9.50	10.11	9.23	8.15
Gross interest expense	5.50	6.78	8.75	8.02	6.36	6.95	5.98	5.02
Net interest margin	3.12	3.09	3.07	3.17	3.15	3.16	3.25	3.13
Noninterest income	0.78	0.89	0.99	1.05	1.12	1.27	1.39	1.46
Loan and lease loss provision	0.24	0.25	0.26	0.39	0.47	0.55	0.66	0.76
Securities gains[2]						-0.01	0.06	0.13
Noninterest expense	2.54	2.63	2.76	2.91	2.95	3.05	3.15	3.17
Income before tax	1.12	1.10	1.04	0.91	0.84	0.82	0.89	0.81
Taxes	0.28	0.28	0.24	0.17	0.18	0.19	0.21	0.19
Other[3]	-0.04	-0.03	-0.04	-0.03	0.00	0.01	0.01	0.01
Return on assets[4]	0.80	0.79	0.76	0.71	0.67	0.64	0.70	0.63
Cash dividends declared	0.28	0.29	0.30	0.31	0.33	0.31	0.33	0.33
Net retained earnings	0.52	0.50	0.46	0.40	0.34	0.33	0.37	0.31
Return on equity[5]	13.90	13.70	13.20	12.20	11.24	10.63	11.33	10.22
Average assets ($ billions)	1,593	1,768	1,940	2,100	2,253	2,398	2,604	2,799

Note: Discrepancies due to rounding error.

[1] Average assets are based on fully consolidated volumes outstanding at the beginning and at the end of the year.

[2] Banks were required to report securities gains or losses above the tax line on their income statements for the first time in 1984.

[3] Includes securities and extraordinary gains or losses after taxes, for 1979-83 data, and extraordinary items and other adjustments after taxes for 1984-86 data.

[4] Return on assets is net income divided by average assets.

[5] Return on equity is net income divided by average equity. Average equity is based on fully consolidated volumes outstanding at the beginning and at the end of the year.

Sources: *Federal Reserve Bulletin,* 1981, 1984 (1979-83 data); Consolidated Reports of Condition and Income (1984-86 data).