John Maynard Keynes

Milton Friedman

John Maynard Keynes (1883–1946) is the latest in a line of great British economists who had a profound influence on the discipline of economics. By common consent, the line starts with Adam Smith (1723–1790), whose *Wealth of Nations* (1776) is generally regarded as the founding document of modern economics. It continues with David Ricardo (1772–1823), whose *Principles of Political Economy* (1817) dominated classical economics for much of the nineteenth century, and, incidentally, provided Karl Marx with one of his central concepts: the labor theory of value. John Stuart Mill's (1806–1873) *Principles of Political Economy*, published in the same year, 1848, as the *Communist Manifesto* by Marx and Engels, became the standard textbook in the English-speaking world—and beyond—for decades. William Stanley Jevons's (1835–1882) *Theory of Political Economy* (1871) inaugurated the "marginal revolution," which replaced, or supplemented, emphasis on cost of production (supply) as determining value with emphasis on utility (demand). He resolved the classic diamond-water paradox—diamonds are a luxury, water a necessity, yet diamonds command a higher price than water—by showing that "marginal utility"—the utility gained from having one more unit of something—not "total utility" plays the key role in determining price. Alfred Marshall (1842–1924), Keynes's own teacher, guide, and patron, dominated economics in the English-speaking world from the publication of the first edition of his classic, *Principles of Economics* (1890), to the 1930s.

The Federal Reserve Bank of Richmond is indebted to Professor Friedman for his kind permission to publish this article in its original English version. The article first appeared in German translation in the volume of commentaries accompanying the facsimile edition of John Maynard Keynes's the *General Theory of Employment, Interest and Money* (1936) published in 1989 by Verlag Wirtschaft und Finanzen GmbH, Düsseldorf, as part of its series *Klassiker der Nationalökonomie.* The generous permission of that publisher and its principal, Mr. Michael Tochtermann, are gratefully acknowledged.

Keynes clearly belongs in this line. In listing "the" classic of each of these great economists, historians will cite the *General Theory* as Keynes's path-breaking contribution. Yet, in my opinion, Keynes would belong in this line even if the *General Theory* had never been published. Indeed, I am one of a small minority of professional economists who regard his *Tract on Monetary Reform* (1923), not the *General Theory*, as his best book in economics. Even after sixty-five years, it is not only well worth reading but continues to have a major influence on economic policy.

1. KEYNES'S LIFE

From 1908 to his death in 1946, Keynes was an active Fellow of King's College, Cambridge, influencing successive generations of students. For many years, he was also Bursar of King's College, and is credited with making it one of the wealthiest of the Cambridge colleges. From 1911 to 1944, he was the editor or joint-editor of the *Economic Journal*, at the time the leading professional economic periodical in the English-speaking world. Simultaneously, he was also Secretary of the Royal Economic Society.

Despite his lifelong commitment to economics, the earliest work he completed—though not the earliest to be published—was in mathematics not economics—*A Treatise on Probability*—essentially completed by 1911, but first published in 1921. It is a mark of Keynes's range, creative originality, and insight that much recent work in statistics has returned to the themes of the *Treatise on Probability*. In economics, his first major publication was *Indian Currency and Finance* (1913), a product of his service in the India Office of the British government from 1906 to 1908.

Monetary Reform (1923) was followed in 1930 by the two-volume *Treatise on Money*, much of which remains of value, though Keynes himself came to regard its theoretical analysis as simply a step on the road to the *General Theory*, the last of his major works. These major works were supplemented by numerous articles, reviews, and biographical essays on some of his predecessors.[1]

Keynes's interest and influence were by no means limited to the confines of the academy. For decades he exerted a major influence on public affairs and played an active role in the world of business. His *Economic Consequences of the Peace* (1919), based on his activities as an adviser to the British Treasury during the negotiation of the Versailles Peace Treaty, had a major impact on public opinion and public policy, not only in Britain but throughout the world, and not only immediately. It was translated into many languages, became a worldwide best-seller, and first established Keynes as a major public

[1] The biographical essays on economists are gathered together in his *Essays in Biography* (1933), along with similar essays on politicians and others.

figure. It influenced the reaction of both victors and vanquished to the Versailles Peace Treaty. Indeed, in a book, *The Carthaginian Peace; or, the Economic Consequences of Mr. Keynes*, published more than two decades later (1946), Etienne Mantoux pays the *Economic Consequences* a backhanded compliment by arguing that Keynes's debunking of the peacemakers was the source of all subsequent evil, including World War II.

From 1919 on, Keynes remained active in public matters, publishing a steady stream of articles on current affairs in nonprofessional journals and newspapers, advising and participating in the deliberations of the Liberal party, serving as chairman of the *Nation and Athenaeum* when it was acquired by a group of Liberals in 1922, and later as director of the combined *New Statesman and Nation*, leading journals of opinion for which he wrote frequently. He brought together many of his most significant pieces on public affairs in *Essays in Persuasion* (1931). He served on government commissions, notably the Macmillan Commission, and advised and consulted with successive governmental ministers. He was chairman of the National Mutual Insurance Company and director of several other insurance companies. His interests were truly catholic: E. A. G. Robinson, who was co-editor of the *Economic Journal* with Keynes for some years and succeeded him as editor, begins an *Encyclopaedia Britannica* article on Keynes by describing him as "1st Baron . . . , British economist who revolutionized economic theories, critic and architect of national economic policies, political essayist, successful financier, bibliophile and patron of the arts." His interest in one particular art, ballet, was both cause and effect of his marriage in 1925 to Lydia Lopokova, a famous Russian ballerina. He established and largely financed the Cambridge Arts Theater and was a trustee of the National Gallery.

From 1919 to World War II, Keynes's connection with government was primarily as an influential outsider. From 1940 on, he served in government in a variety of capacities concerned with the economic conduct of the war and postwar reconstruction. He was the chief British representative at Bretton Woods in 1944, where he was a major architect of the plans for the International Monetary Fund and the World Bank for Reconstruction and Development. He was the chief negotiator of the large U.S. loan to Britain in 1945. On his return to Britain, he played an important role in persuading the British Parliament to adopt the Bretton Woods agreement. He died shortly thereafter, on April 21, 1946.

2. THE INFLUENCE OF THE *GENERAL THEORY*

To return to the *General Theory*: its influence on both economic thinking and economic practice was profound. The "Keynesian revolution" was far more than a figure of speech. From shortly after the publication of the book in 1936 to at least the 1960s, the majority of professional economists, and certainly

the most prominent, termed themselves "Keynesians." Those who called themselves non- or anti-Keynesians were a beleaguered minority, supplemented, it must be said, by some important writers on economics who were not members of the professional guild.[2] Governments around the world hastened to adopt "Keynesian policies," though many an economist—both Keynesians and anti-Keynesians—regarded some of the policies, particularly when they led to inflation, as at best "bastard Keynesianism."[3]

As of this writing (1988), the status and influence of the book has changed. It continues to have a major influence on economic thinking and economic policy, and will long continue to do so, but for very different reasons and in a very different way than it did initially. The catalyst for the change was the inflation and stagflation of the 1970s. As Robert Lucas wrote in 1981, "Proponents of a class of models which promised $3\frac{1}{2}$ to $4\frac{1}{2}$ percent unemployment to a society willing to tolerate annual inflation rates of 4 to 5 percent have some explaining to do after a decade such as we have gone through [i.e., the 1970s, when inflation rose to 16 percent and unemployment to 8 percent in the United States, and to 30 percent and 6 percent in the U.K. Inflation rose as high as 25 percent in Japan and 7 percent in Germany, though unemployment remained relatively low]. A forecast error of this magnitude and central importance to policy has consequences, as well it should."

The predictions to which Lucas refers were based on the so-called Phillips curve which linked inflation inversely to unemployment—allegedly, the higher the rate of inflation, the lower the level of unemployment. The curve was asserted by many Keynesians to be stable over time and to specify a menu of combinations of inflation and unemployment, any of which was attainable by the appropriate monetary and fiscal policy. Lucas went on to note that "in the late 1960s Milton Friedman (1968) and Edmund Phelps (1968) had argued . . . that these predicted Phillips curve trade-offs were spurious." They emphasized the importance of distinguishing between anticipated and unanticipated inflation in interpreting the Phillips curve, and Friedman introduced the concept of a "natural rate of employment" to which the economy would tend as economic actors adjusted their anticipations.

"The central forecast to which [Friedman's and Phelps's] reasoning led," Lucas continued, "was a conditional one, to the effect that a high-inflation decade should not have less unemployment on average than a low-inflation decade. We got the high-inflation decade, and with it as clear-cut an

[2] In the U.S., the most important was doubtless Henry Hazlitt, *The Failure of the New Economics: An Analysis of the Keynesian Fallacies* (Princeton, N.J.: Van Nostrand, 1959).

[3] The phrase was coined by Joan Robinson, one of the earliest and most dedicated members of Keynes's inner circle, in her review of Harry Johnson's *Money, Trade and Economic Growth* (1962), *Economic Journal*, vol. 72 (September 1962), p. 690. However, she used it to refer to the theories of some of Keynes's followers, rather than to policies.

experimental discrimination as macro-economics is ever likely to see, and Friedman and Phelps were right."[4]

The 1980s have been no kinder to the earlier Keynesian models. In the U.S., inflation was brought down drastically, accompanied by a temporary increase in unemployment to a peak of nearly 11 percent—a short-term reaction to unanticipated disinflation along Phillips curve lines. But then, from 1983 on, unemployment fell concurrently with further declines in inflation, reaching 6 percent by the end of 1987 when inflation was about 4 percent—a flat contradiction of the asserted negative relation between unemployment and inflation embodied in the Phillips curve. In the U.K., too, an initial decline in inflation was accompanied by a sharp rise in unemployment, which was very much slower to decline but has more recently begun to do so. In Germany, inflation has come down since the early 1980s; unemployment rose initially, as in the U.S. and the U.K., but, in contrast to them, continued to rise after inflation had settled down, and has remained high. Japan, which was the first of the major countries to cut sharply the rate of inflation, has succeeded in keeping inflation low with little change in its recorded unemployment rate. All in all, this experience is hardly consistent with a stable trade-off between inflation and unemployment.

Experience led to disillusionment with initial Keynesianism on the part not only of professional economists but also of policymakers. The most dramatic evidence came from James Callaghan, when he was the Labour prime minister of the U.K.—the party and the country that had gone farthest in embracing and adopting Keynesian policies. Said Callaghan in 1976, "We used to think that you could just spend your way out of a recession and increase employment by cutting taxes and boosting government spending. I tell you, in all candour, that that option no longer exists; and that insofar as it ever did exist, it only worked by injecting bigger doses of inflation into the economy followed by higher levels of unemployment as the next step. That is the history of the past twenty years."

Despite the widespread rejection of some of the key propositions that constituted the "Keynesian revolution," the book continues to have a major impact on economic thinking. Some indication of its influence is given by the continuing citations to the book in the professional literature. Data from one citation index, which covers a wide range of economic journals, are available for sixteen years, 1972 to 1987. In all, there were 1,558 citations to the *General Theory,* or an average of nearly 100 a year. Of the total, 729 occurred in the first eight years, 829 in the second eight, so there is no sign that interest in the book is declining. However, the character of the book's influence has changed.

[4] Robert E. Lucas, Jr., "Tobin and Monetarism: A Review Article," *Journal of Economic Literature,* vol. 19 (June 1981), p. 560.

Some years ago, I remarked to a journalist from *Time* magazine, "We are all Keynesians now; no one is any longer a Keynesian." In regrettable journalist fashion, *Time* quoted the first half of what I still believe to be the truth, omitting the second half. We all use Keynesian terminology; we all use many of the analytical details of the *General Theory*; we all accept at least a large part of the changed agenda for analysis and research that the *General Theory* introduced. However, no one accepts the basic substantive conclusions of the book, no one regards its implicit separation of nominal from real magnitudes as possible or desirable, even as an analytical first approximation, or its analytical core as providing a true "general theory."

As one, no doubt somewhat idiosyncratic, view of the book, I quote from a reply that I wrote some years ago to criticisms of my work mostly from a "Keynesian" point of view:

> "One reward from writing this reply has been the necessity of rereading earlier work, in particular [Keynes's] . . . *General Theory*. The *General Theory* is a great book, at once more naive and more profound than the 'Keynesian economics' that Leijonhufvud contrasts with the 'economics of Keynes.' . . .[5]
>
> "I believe that Keynes's theory is the right kind of theory in its simplicity, its concentration on a few key magnitudes, its potential fruitfulness. I have been led to reject it, not on these grounds, but because I believe that it has been contradicted by evidence: its predictions have not been confirmed by experience. This failure suggests that it has not isolated what are 'really' the key factors in short-run economic change.
>
> "The *General Theory* is profound in the wide range of problems to which Keynes applies his hypothesis, in the interpretations of the operation of modern economies and, particularly, of capital markets that are strewn throughout the book, and in the shrewd and incisive comments on the theories of his predecessors. These clothe the bare bones of his theory with an economic understanding that is the true mark of his greatness.
>
> "Rereading the *General Theory* has . . . reminded me what a great economist Keynes was and how much more I sympathize with his approach and aims than with those of many of his followers."[6]

3. THE MESSAGE OF THE *GENERAL THEORY*

As its title indicates, the *General Theory* is almost pure abstract theory. There is only passing reference to applied economics, statistical magnitudes, or economic policy. Yet, like all of Keynes's writings on economics, it was inspired by a major contemporary problem and written in the hope and expectation

[5] Axel Leijonhufvud, *On Keynesian Economics and the Economics of Keynes* (London: Oxford University Press, 1968).

[6] *Milton Friedman's Monetary Framework: A Debate with His Critics*, ed. by Robert J. Gordon (Chicago: University of Chicago Press, 1974), pp. 133–34.

of providing a solution. The book was written during the worldwide Great Depression following 1929, when idle men, idle machines, and unmet demand coexisted on a large scale for years on end and produced widespread poverty, misery, and deprivation. For Britain, it followed a near-decade of economic stagnation, high unemployment, and long-term dependence of many families on a government dole. The key problem of the time was how to explain the apparent paradox, and, more urgently, how to resolve it.

Ups and downs in economic activity involving occasional periods of widespread unemployment had long occurred and had engaged the attention of numerous economists under the rubric of "business fluctuations," or "business cycles." Various theories had been offered to explain them. Most earlier theories implicitly accepted the proposition that a private-enterprise capitalist system contained self-correcting forces that would keep disturbances temporary. By corrective adjustments to changes in circumstances, the system, it was believed, would tend toward full employment of both men and machines—save only for fractional and transitory unemployment implicit in a dynamic economy. However, the long duration and magnitude of the unemployment during the Great Depression and the prior years in Britain did not seem to fit this pattern. Could these be interpreted as simply a temporary, if long-lasting, disturbance? Or did they indicate a defect in the supposed self-adjusting forces at work, so that the economy could get stuck for long periods of time at a position of high unemployment—a position that might have just as much reason to be regarded as an "equilibrium" as a position of full employment?

Such a possibility had frequently been asserted by socialist and other critics of a capitalist system, whom the mainline professional economists had regarded as "crackpots." Keynes took the possibility seriously and proceeded to construct an hypothesis that he believed demonstrated the possibility—indeed the frequent reality—that, without government intervention, a private-enterprise capitalist system using a non-commodity money would tend toward a position characterized by a high level of involuntary unemployment of persons who would willingly be employed at the current wage rate but could not find jobs.

The classical remedy for idle men, according to Keynes, was a decline in the real wage rate, which would reduce the number of persons seeking jobs and increase the number of persons employers wanted to hire. The classical remedy for idle machines was a reduction in the cost to enterprises of using and producing such machines, and that was expected to occur via a reduction in the real interest rate.

In the 1920s and 1930s in Britain, these classical remedies seemed either inoperative or ineffective. Keynes set himself the task of explaining why, of constructing an alternative theory that would both explain what was happening and justify alternative policies—such as the large public works programs he had been recommending since the mid-1920s.

In one sense, his approach was strictly Marshallian: in terms of demand and supply. However, whereas Marshall dealt with specific commodities and "partial equilibrium," Keynes proposed to deal with what he called "aggregate demand" and the "aggregate supply function," and with general not partial equilibrium.

Where he deviated from Marshall was in the key variables that he regarded as producing equilibrium between demand and supply and in the process of adjustment to a change in demand or supply. In Marshallian analysis, the key role was played by prices, which reacted quickly to any change in circumstances. Let there be a sudden increase in demand, in the sense of a demand function relating the quantity demanded to price. In Marshall's view, the immediate reaction would be on prices, which would rise to choke off the quantity demanded to the prior level plus whatever additional quantities might be made available from inventories. The rise in prices during the "market period" would give producers an incentive to increase output in the "short run" by using existing plant and equipment more intensively, and, if the increased demand persisted, in the longer run by adding to plant and equipment. In short, prices adjusted rapidly, quantities slowly, and changes in prices played the major role in producing equilibrium.

To Keynes, it seemed clear that this process had been inoperative or ineffective with respect to the economy as a whole. Nominal wage rates had indeed declined, but so had nominal prices, so that real wages had hardly moved, and may indeed have increased. He concluded that movements in prices and interest rates could not be counted on. Accordingly, he reversed Marshall's presumptions: prices of labor and capital, at least "real wages" and "real interest rates," are very slow to adjust; quantities, which is to say consumption, investment, and their sum, total output, are highly flexible and adjust rapidly. Changes in output (aggregate supply), not in prices, play the major role in producing equilibrium. Accordingly, as a first approximation—though one he never really relaxed—he took prices as given by forces outside his analysis. As a first approximation, also, he abstracted from both government spending and international trade, but these could readily be integrated into the analysis without affecting its substance.

Keynes defined aggregate demand and aggregate supply in terms of employment, in line with his view that he was developing a "theory of employment." However, both Keynes and his followers tended to replace employment by output and to express aggregate demand and aggregate supply in terms of the value of output demanded by the public and supplied by enterprises.

Aggregate demand, in these terms, is the sum of expenditures on consumption goods and expenditures on investment goods. Keynes regarded expenditures on consumption as depending on income, introducing one of his key concepts: the propensity to consume, or, in his words, "the functional relationship . . . between . . . a given level of income in terms of wage-

units, and . . . the expenditure on consumption out of that level of income." A "fundamental psychological law," which plays a key role in the Keynesian system, is that "men are disposed . . . to increase their consumption as their income increases, but not by as much as the increase in their income"—i.e., the "marginal propensity to consume" is less than unity.[7]

Keynes defined investment as "the current addition to the value of the capital equipment which has resulted from the productive activity of the period." He regarded investment as depending on the "marginal efficiency of capital," the second of his key concepts, which he defined as "that rate of discount which would make the present value of the series of annuities given by the returns expected from the capital-asset during its life just equal to its supply price," i.e., "the cost of producing" one more unit of the asset. Like the propensity to consume, the marginal efficiency of capital is a function or schedule relating the amount of investment to the interest rate, since entrepreneurs would have an incentive to add to investment so long as the yield exceeded the interest rate at which they could borrow the funds to finance the investment.[8]

The interest rate, in turn, he regarded as determined by "liquidity preference," the third of his key concepts. "An individual's liquidity-preference is given by a schedule of the amounts of his resources, valued in terms of money or of wage-units, which he will wish to retain in the form of money in different sets of circumstances." He regarded the amount of their assets that individuals would want to hold in the form of money as depending on both income and the interest rate—income because that would affect the amount held for "transactions- and precautionary-motives," the interest rate, because that would affect the amount held "to satisfy the speculative-motive."[9]

If, as Keynes did, we let Y be income, identical with the value of output, C be consumption, I be investment, L liquidity preference, M the quantity of money, and r the interest rate, then aggregate demand is given by

$$Y = C(Y) + I(r), \tag{1}$$

and the demand for money by

$$M = L(Y, r). \tag{2}$$

In line with his implicit assumption about the relative speed of adjustment of prices and output, Keynes regarded supply as essentially passive, expanding or contracting as demand expanded or contracted, subject only to the proviso that employment is less than "full," which he defined as the point at which an increase in aggregate demand would call forth no additional workers willing to

[7] *The General Theory of Employment Interest and Money* (London: Macmillan, 1936), pp. 90, 96, and 114.

[8] Ibid., pp. 62 and 135.

[9] Ibid., pp. 166 and 199.

work at the wage offered. This leads him to regard aggregate supply as given simply by aggregate demand, or

$$Y_S = Y_D, \tag{3}$$

and the level of aggregate supply and demand as affecting not a price but solely employment.

If we regard the interest rate as fixed, along with other prices, then equations (1) and (3) define the famous Keynesian "multiplier" (attributed by Keynes to Richard Kahn). For a simple version, assume that the consumption function is linear:

$$C = a + bY, \tag{4}$$

with b, of course, less than one. Substituting (4) in (1) and solving for Y, we have

$$Y = \frac{a + I(r)}{1 - b} = \left(\frac{1}{1 - b}\right)[a + I(r)]. \tag{5}$$

The multiplier is $1/(1 - b)$, which, given that b is between zero and unity, is necessarily greater than unity. The multiplicand, $(a + I)$, came to be termed "autonomous" spending, i.e., spending not dependent on the level of income. In addition, once government was introduced into the analysis, autonomous spending was regarded as including not only autonomous consumption spending (a) and investment (I) but also government spending.

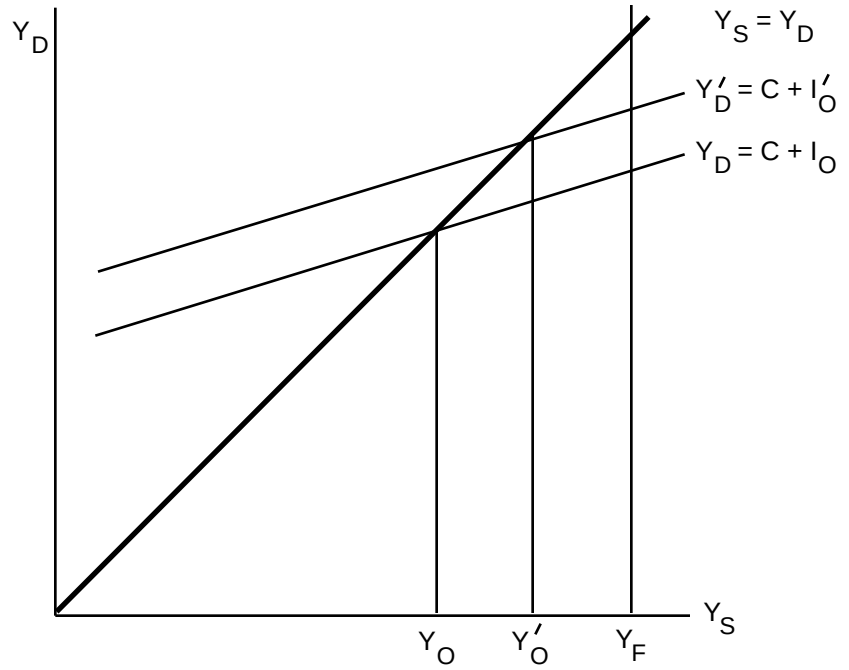
Equations (1) and (3) define also the equally famous "Keynesian cross," which has been reproduced in literally hundreds of textbooks in the past half century and is reproduced here in Figure 1.

The graph makes clear the key importance of the "fundamental psychological law" that the marginal propensity to consume is less than unity. If it were unity, the Y_D line would parallel the Y_S line and there would be either no or an infinite number of equilibrium positions, according as the two parallel lines were distinct or identical. If it exceeded unity, the Y_D line would slope more steeply than the Y_S line, and any point of intersection would be an unstable equilibrium position. Because it slopes less steeply, the intersection at Y_O is a stable equilibrium. If output were temporarily higher than Y_O, employers would be making losses, since the aggregate supply price would exceed aggregate demand, and would seek to contract output. Conversely, if output were temporarily lower than Y_O, employers would be making profits and would seek to expand.

If, for whatever reason, investment were to increase from I_O to I'_O, the Y_D line would shift to Y'_D and the new equilibrium would shift to Y'_O. At Y_F, the point of full employment, the process would end, and "the crude quantity theory of money," which is the particular object of Keynes's scorn and derision—no doubt because of his long earlier adherence to it—"is fully satisfied."[10]

[10] Ibid., p. 289.

Figure 1 The Keynesian Cross



Marvelously simple. A key that apparently unlocks the mystery of long-continued unemployment: inadequate autonomous spending or too low a propensity to consume. Increase either, or both, being careful simply not to go too far, and full employment could be attained. What a wonderful prescription: for consumers, spend more out of your income, and your income will rise; for governments, spend more, and aggregate income will rise by a multiple of your additional spending; tax less, and consumers will spend more with the same result. Though Keynes himself, and even more, his disciples, produced much more sophisticated and subtle versions of the theory, this simple version contains the essence of its great appeal to non-economists and especially governments. Here was one of the most famous and respected economists in the world informing governments that the way to full employment was paved with higher spending and lower taxes. What more attractive advice could politicians wish for? Long regarded public vices turned into public virtues!

Marvelously simple, yes. But also simply marvelous. How could a position such as Y_O in Figure 1 be regarded as a long-term equilibrium—as was implied in the claim that the theory was "general"? At that point, men and machines

are idle. Would not the excess supply of men and machines exert downward pressures on the prices of both? Yes, said Keynes, but, if effective, that would be accompanied by lower money prices of output that would cancel the lower money wages and money cost of capital, so that real wages and the real cost of capital would be unaffected—which is why Keynes expressed all aggregate magnitudes in "wage-units." Hence, said Keynes, flexible wages and prices would do no good. Far better to operate directly on spending.

Of course, Keynes recognized that changes in prices, interest rates, and quantity of money did have effects that provided alternative avenues of escape from the so-called "underemployment equilibrium." At best, it was a transitory equilibrium position, the existence of which would set in motion self-corrective forces. But Keynes tended to rule out these alternative avenues of escape as of no practical significance because of his empirical judgment that prices, wages, and interest rates were highly sluggish. Indeed, some commentators on Keynes maintain that he deliberately overstated his case in order to shock the economics profession into paying attention—a tactic that is common to every innovator, whether it be of an idea or a product.

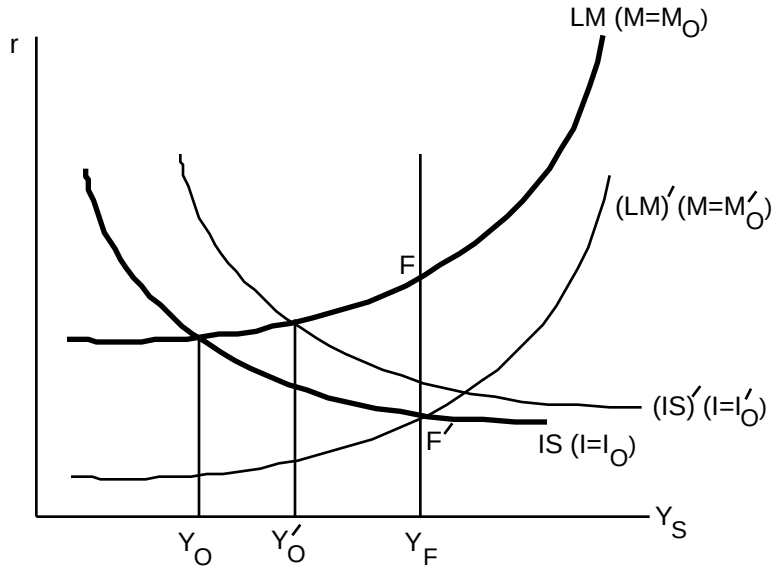
Only one alternative avenue of adjustment is explicitly present in equations (1) and (2)—via the interest rate and the quantity of money. This avenue, analyzed at some length in the *General Theory,* and found wanting to produce, by itself, a full employment equilibrium, also was rapidly incorporated in an alternative, more sophisticated graphical representation of the Keynesian system developed almost simultaneously by John Hicks and Roy Harrod.[11] Figure 2 presents Hicks's IS-LM version, which very quickly became the orthodox version.

In this diagram, the vertical axis is the interest rate. The horizontal axis is income expressed in wage-units, so that it is also output and employment. The IS curve traces equation (5), i.e., it shows the combinations of interest rate and output that would satisfy equation (1): the higher the interest rate, the lower investment and hence income, and conversely, which is why the IS curve has a negative slope. Put differently, it shows the combinations of interest rate and output at which the amount some people wish to invest is equal to the amount other people wish to save, which is what explains the S in IS. But note that the accommodation of saving to investment is produced not by the direct effect of the interest rate on saving, but by the effect of the level of income on saving, via the propensity to consume.

The LM curve traces equation (2) for a fixed quantity of money. Here, the higher the interest rate, the lower the quantity of money that the public would want to hold for a given income, and hence the higher income must be in order

[11] John R. Hicks, "Mr. Keynes and the 'Classics': A Suggested Interpretation," *Econometrica,* vol. 5 (April 1937), pp. 147–59; Roy F. Harrod, "Mr. Keynes and Traditional Theory," *Econometrica,* vol. 5 (January 1937), pp. 74–86.

Figure 2 The IS-LM Diagram



for the actual quantity of money to be willingly held. Hence the positive slope of the LM curve.

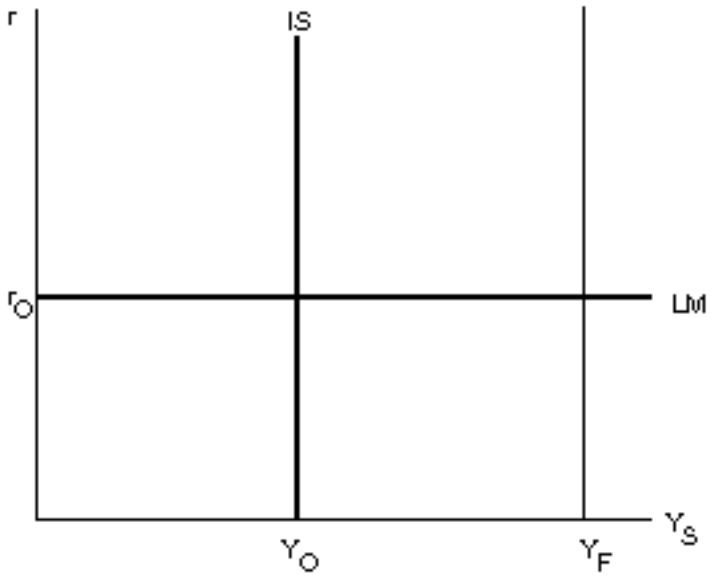
The intersection of the IS and LM curve at Y_O is the counterpart of the intersection of the aggregate demand and supply curves in Figure 1 at Y_O. Similarly, the IS′ curve is the counterpart of the Y'_O curve in Figure 1, reflecting a higher level of investment. It is the IS curve moved to the right by the change in income assumed to be produced by the increase in investment—the change in investment times the investment multiplier.

What is new in Figure 2 are the LM curves. Each LM curve is for a specific quantity of money: the LM curve for $M = M_O$, the (LM)′ curve for $M = M'_O$, which is larger than M_O. For the community to hold the larger quantity of money willingly, either the interest rate must be lower for a given income or income higher for a given interest rate, which is why the (LM)′ curve is to the right of the LM curve.

The IS curve in the diagram embodies a possible Keynesian escape from underemployment via increases in investment (or, more generally, autonomous spending including government spending). Let autonomous spending be high enough so that the IS curve intersects the LM curve at point *F*, and full employment would be attained with the initial quantity of money.

The LM curve offers an alternative escape via the quantity of money. Let the quantity of money be large enough so that the LM curve intersects the

Figure 3 An Extreme IS-LM Diagram with Perfectly Elastic Liquidity Preference and Inelastic Investment

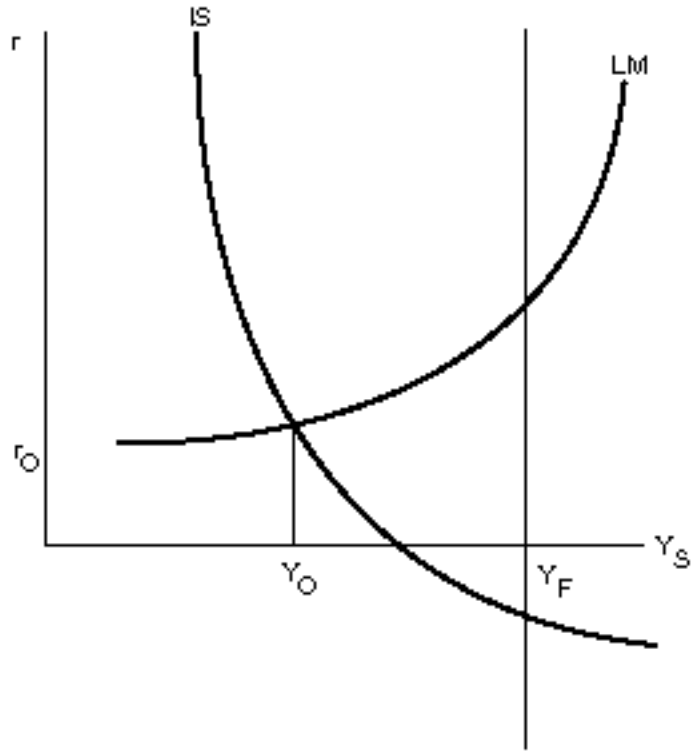


IS curve at point F', and full employment would be attained with the initial marginal efficiency of capital schedule.

Keynes and his followers rejected this possibility as highly unrealistic, largely on the alleged empirical grounds that (1) private autonomous expenditures were little affected by changes in the interest rate while (2) there was a floor to the interest rate at which the community would be willing to hold assets other than money, so that, in the neighborhood of this floor, the quantity of money the community would be willing to hold would be highly sensitive to the interest rate: in short, a low elasticity of investment, but a high elasticity of liquidity preference, with respect to the interest rate.

Figure 3 shows an extreme version of these assumptions: perfectly inelastic investment and perfectly elastic liquidity preference. We are back to the Keynesian cross of Figure 1. No changes in the quantity of money can produce a full employment equilibrium. This LM curve depicts a "liquidity trap," of which Keynes wrote, "whilst the limiting case might become practically important in future, I know of no example of it hitherto. Indeed, owing to the unwillingness of most monetary authorities to deal boldly in debts of long term, there has

Figure 4 A Less-Extreme Liquidity Trap



not been much opportunity for a test."[12] Of course, it is not necessary to go to this extreme to generate Keynesian unemployment equilibria, and Keynes and his followers did not, though some of the more enthusiastic of his disciples came very close during the high tide of the Keynesian revolution. It is only necessary to suppose a highly inelastic IS curve, and a highly elastic LM curve, as in Figure 4. In this version, a negative interest rate would be required for a full employment equilibrium. The Keynesians ruled out this possibility by the assumption of given prices.[13]

The avenue of adjustment that is not explicitly allowed for in either equations (1) and (2) or in the more sophisticated IS-LM diagram is the level of prices and wages. As already noted, a Keynesian position of underemployment equilibrium means downward pressure on wages and prices. Keynes explicitly

[12] *The General Theory,* p. 207.

[13] The interest rate that is relevant to investment is the "real" interest rate, i.e., the nominal rate of interest less the rate of inflation, and the "real" interest rate has often been negative.

recognized that a change in real wages would affect employment by altering both the supply and the demand for labor.[14] However, he ruled out that avenue of escape on the grounds that prices and wages would tend to change pari-passu leaving real wages largely unchanged—not a bad empirical approximation for the kind of major disturbances, such as the Great Depression, whose origin and cure Keynes was seeking. Keynes discussed two other effects of changes in the level of prices and wages. The first is on the real quantity of money, and thence the rate of interest. A lower level of prices is equivalent to a higher quantity of money, and like an increase in the quantity of money would shift the LM curve to the right. The second is the effect of a lower rate of interest on the consumption function, an effect that has come to be called the Keynes effect. The lower the interest rate, the higher the capital value of a given stream of income—such as rent on a piece of land, or coupons on a bond. Hence, a lower interest rate increases the wealth of the community. The higher the wealth, the less pressure to add to wealth via savings, and hence the higher is likely to be the average and marginal propensity to consume at any income.

Though Keynes recognized the existence of these avenues of adjustment, he largely dismissed them on empirical grounds. Sluggishness of price movements had pride of place, but inelasticity of investment and elasticity of liquidity preference with respect to the interest rate and inelasticity of consumption with respect to wealth were also important.

A third effect of a pari-passu change in prices and wages, which came to be known as the "Pigou" effect, was not discussed explicitly by Keynes. The lower the price level, the higher the real value of the fixed quantity of money. In principle, there is no limit to the real value of a fixed nominal quantity of money, and hence no limit to the wealth of a community, and accordingly, no limit to the extent to which the IS curve could be shifted to the right by the reduction in the incentive to save.[15] There is much dispute about the empirical importance of this effect. I personally regard it as minor. However, on the purely abstract theoretical level of the *General Theory,* it conclusively demonstrates that there is no such flaw in the price system as Keynes professed to demonstrate. His position of underemployment equilibrium, whatever else it might be, was not a long-run equilibrium position that set in motion no effective forces tending toward full employment.

What difference does this abstract analysis make? Is it not simply arguing about how many angels can dance on the point of a pin? The answer is that it destroys Keynes's most striking and radical claim made in the first paragraph of the *General Theory*: that what he called the "classical economics," and, in

[14] See, for example, ibid., p. 289.

[15] For a fuller theoretical analysis of (a) the possibility of a negative equilibrium interest rate, and (b) the Keynes and Pigou effects, see Milton Friedman, *Price Theory* (Chicago: Aldine Publishing Co., 1976), pp. 313–21.

particular, the quantity theory of money, were fundamentally fallacious, "that the postulates of the classical theory are applicable to a special case only and not to the general case, the situation which it assumes being a limiting part of the possible positions of equilibrium. Moreover, the characteristics of the special case assumed by the classical theory happen not to be those of the economic society in which we actually live, with the result that its teaching is misleading and disastrous if we attempt to apply it to the facts of experience."[16]

If this extreme claim is wrong, Keynes's theory becomes not a theory of "equilibrium" but at best a theory of disequilibrium, readily encompassed in the earlier orthodoxy. Conventional wisdom prior to the *General Theory* had always recognized that fluctuations existed, and that periods of widespread unemployment did occur from time to time. But it regarded these as responses to changes in circumstances, plus rigidities in prices, wages, and other variables that impeded rapid adjustment to the new circumstances. And, indeed, conventional economic wisdom has by now come to regard the Keynesian theory as a theory of disequilibrium, which provides a useful way to analyze the process of adjustment to changes in circumstances in a world of relatively rigid prices and wages. It should be added that there does remain a significant number of respected economists who continue to regard Keynes's contribution as providing a truly general theory fully justifying his initial claims, and continue to regard him as having demolished the so-called classical theory.[17]

There remains the twin questions of why Keynes, who described himself in the preface to the German edition as having been "a priest of" the English classical quantity theory tradition, regarded it as incompetent to explain the persistence of high unemployment in the 1920s and 1930s, and of how those of us who disagree with him reconcile that remarkable phenomenon with the earlier theory. The key to the answer to both questions is the interpretation of monetary developments, and particularly monetary policy in the 1930s. Consider first the situation in the U.S. By contrast with Britain, the 1920s were a period of general prosperity, high employment, and relatively stable prices. There was no reason to question the importance of monetary policy. Indeed, the Federal Reserve System in the United States took for itself much of the credit for the good performance of the economy. But then came the Great Depression. Its initial phase, from 1929 to late 1930, had all the characteristics of a garden-variety recession, though somewhat more severe than most, and, indeed, had it ended in early 1930, or even early 1931, as it showed some signs of doing, it would have gone down in history in that way, not as a major contraction, let alone Great Depression. But the second phase, from the end of 1930 to 1933, was very different. It was marked by a succession of banking crises, and the

[16] Ibid., p. 1.

[17] The most prominent of this group are the late Joan Robinson, the late Nicholas Kaldor, in Britain, and Professor Robert Eisner, in the United States.

veritable collapse of the banking system leading to an unprecedented "bank holiday" in March 1933, during which all the banks of the country—including the Federal Reserve Banks themselves—were closed for business. When the holiday ended and "sound banks" reopened, they numbered only two-thirds as many as were in existence in 1929. This sequence of events was accompanied by a disastrous increase in unemployment, and major declines in prices, wages, and national income both in current and constant prices. From 1929 to 1933, "money income fell 53 percent and real income 36 percent . . . . Per capita real income in 1933 was almost the same as in the depression year of 1908, a quarter of a century earlier . . . . At the trough of the depression one person was unemployed for every three employed."[18] And what happened in the United States was duplicated—the banking disaster partly excepted—around the world.

To Keynes and many of his contemporaries, this sequence of events seemed a clear contradiction of the earlier theory and of the efficacy of monetary policy. They tended then, as many still do, to regard monetary policy as operating via interest rates. Short-term interest rates in the United States had fallen drastically during the contraction. In particular, the discount rate charged by the Federal Reserve Banks on loans to banks that were members of the Federal Reserve System was steadily reduced from 6 percent in 1929 to 1.5 percent by the fall of 1931, though it was then abruptly increased to 3.5 percent in response to Britain's departure from gold in September 1931, and was still 2.5 percent in early 1933. Judged in these terms, monetary policy was "easy," yet it apparently had been powerless to stem the contraction, giving rise to widespread apprehension that monetary policy was like a string: you could pull on it, but not push on it, i.e., monetary policy could check inflation but could not offset contraction.

From another, and I would argue far more significant, point of view, monetary policy was anything but "easy." That point of view regards monetary policy as operating via the quantity of money. In terms of annual averages, the quantity of money in the United States fell by one-third from 1929 to 1933—by 2 percent from 1929 to 1930, just before the onset of the first banking crisis, and by a further 32 percent from 1930 to 1933. Data on the quantity of money were not published regularly at that time and were not readily available even with some lag, whereas interest rates were readily and contemporarily available—both effect and reinforcement of the tendency to interpret monetary policy in terms of the interest rate rather than the quantity of money.

Keynes may well not have known what was happening to the quantity of money, though if he had, he would also have known that "[a]t all times

[18] Milton Friedman and Anna J. Schwartz, *A Monetary History of the United States, 1867–1960* (Princeton: Princeton University Press for the National Bureau of Economic Research, 1963), p. 301.

throughout the 1929–33 contraction, alternative policies were available to the [Federal Reserve] System by which it could have kept the stock of money from falling, and indeed could have increased it at almost any desired rate." Far from demonstrating, as Keynes concluded, that monetary policy is impotent, "[t]he contraction is in fact a tragic testimonial to the importance of monetary forces."[19] The contraction continued and deepened not because there were no equilibrating forces within the economy but because the economy was subjected to a series of shocks succeeding one another: a first banking crisis beginning in the fall of 1930, a second beginning in the spring of 1931, Britain's departure from gold in September 1931, and the final banking crisis beginning in January 1933—all accompanied by a decline in the quantity of money of 7 percent from 1930 to 1931, 17 percent from 1931 to 1932, and 12 percent from 1932 to 1933.

Even after the end of the contraction and the start of revival in 1933, the shocks continued and impeded recovery: major legislative measures during Franklin Delano Roosevelt's New Deal that interfered with market adjustments and generated uncertainty within the business community, although some of them, particularly the enactment of federal insurance of bank deposits, reassured the community about the safety and stability of the financial institutions; then ill-advised monetary measures in 1936 that halted the rapid rise that had been occurring in the quantity of money and produced an absolute decline from early 1937 to early 1938 that exacerbated if it did not produce the accompanying severe cyclical decline.

Keynes's readiness to interpret the U.S. experience as evidence of the impotence of monetary policy was greatly strengthened by the British experience. By contrast with the U.S., the 1920s was a period of stagnation and high unemployment that the severe worldwide contraction beginning in 1929 intensified. However, the contraction ended earlier in Britain than in the U.S., shortly after Britain left the gold standard and thereby cut its monetary link with the U.S. Here, too, a succession of shocks played an important role: the end of World War I and demobilization; the pressure to return to gold at the prewar parity, which required internal deflation; the return in 1925 to gold at a parity that overvalued the pound sterling, particularly after France returned to gold at a parity that undervalued the franc; and, finally, the shock waves that spread from the U.S. after 1929. The effect of steady deflationary pressure was reinforced by "an unemployment insurance scheme that paid benefits that were high relative to wages available subject to few restrictions . . . . Although a few interwar observers saw clearly the effects of unemployment insurance, Keynes and his followers did not."[20]

---

[19] Ibid., pp. 693 and 300.

[20] Daniel K. Benjamin and Levis A. Kochin, "Searching for an Explanation of Unemployment in Interwar Britain," *Journal of Political Economy*, vol. 87 (June 1979), p. 441.

4. KEYNES'S POLITICAL INFLUENCE

In judging Keynes's overall influence on public policy, it is necessary to distinguish his bequest to technical economics from his bequest to politics. Keynes's bequest to technical economics was strongly positive. His bequest to politics, in my opinion, was not. Yet I conjecture that his bequest to politics has had far more influence on the shape of today's world than his bequest to technical economics. In particular, it has contributed greatly to the proliferation of overgrown governments increasingly concerned with every phase of their citizens' daily lives.[21]

I can best indicate what I regard to be Keynes's bequest to politics by quoting from his famous letter to Professor Friedrich von Hayek praising Hayek's *Road to Serfdom*. The part generally quoted is from the opening paragraph of the letter: "In my opinion it is a grand book . . . . [M]orally and philosophically I find myself in agreement with virtually the whole of it; and not only in agreement with it, but in a deeply moved agreement."

The part I want to direct attention to comes later:

> "I should therefore conclude your theme rather differently. I should say that what we want is not no planning, or even less planning, indeed I should say that we almost certainly want more. But the planning should take place in a community in which as many people as possible, both leaders and followers wholly share your own moral position. Moderate planning will be safe if those carrying it out are rightly orientated in their own minds and hearts to the moral issue.
>
> "What we need therefore, in my opinion, is not a change in our economic programmes, which would only lead in practice to disillusion with the results of your philosophy; but perhaps even the contrary, namely, an enlargement of them . . . . No, what we need is the restoration of right moral thinking—a return to proper moral values in our social philosophy . . . . Dangerous acts can be done safely in a community which thinks and feels rightly, which would be the way to hell if they were executed by those who think and feel wrongly."[22]

Keynes was exceedingly effective in persuading a broad group—economists, policymakers, government officials, and interested citizens—of the two concepts implicit in his letter to Hayek: first, the public interest concept of government; second, the benevolent dictatorship concept that all will be well if only good men are in power. Clearly, Keynes's agreement with "virtually the

---

[21] The rest of this preface up to the final paragraph is drawn largely from my "Comment on Leland Yeager's Paper on the Keynesian Heritage," in *The Keynesian Heritage*, a symposium by Leland Yeager, Milton Friedman, and Karl Brunner, Center Symposia Series CS–16 (Rochester, N.Y.: Center for Research in Government Policy and Business, Graduate School of Management, University of Rochester, 1985), pp. 12–18.

[22] Donald Moggridge, ed., *John Maynard Keynes, The Collected Writings*, Vol. XXVII: *Activities, 1940–1946*, pp. 385, 387, 388.

whole" of the *Road to Serfdom* did not extend to the chapter titled "Why the Worst Get on Top."

Keynes believed that economists (and others) could best contribute to the improvement of society by investigating how to manipulate the levers actually or potentially under control of the political authorities so as to achieve desirable ends, and then persuading benevolent civil servants and elected officials to follow their advice. The role of voters is to elect persons with the right moral values to office and then let them run the country.

From an alternative point of view, economists (and others) can best contribute to the improvement of society by investigating the framework of political institutions that will best assure that an individual government employee or elected official who, in Adam Smith's words, "intends only his own gain . . . is . . . led by an invisible hand to promote an end that was no part of his intention," and then persuading the voters that it is in their self-interest to adopt such a framework. The task, that is, is to do for the political market what Adam Smith so largely did for the economic market.

Keynes's view has been enormously influential—if only by strongly reinforcing a pre-existing attitude. Many economists have devoted their efforts to social engineering of precisely the kind that Keynes engaged in and advised others to engage in. And it is far from clear that they have been wrong to do so. We must act within the system as it is. We may regret that government has the powers it does; we may try our best as citizens to persuade our fellow citizens to eliminate many of those powers; but so long as they exist, it is often, though by no means always, better that they be exercised efficiently than inefficiently. Moreover, given that the system is what it is, it is entirely proper for individuals to conform and promote their interests within it.

An approach that takes for granted that government employees and officials are acting as benevolent dictators to promote in a disinterested way what they regard as the public's conception of the "general interest" is bound to contribute to an expansion in governmental intervention in the economy—regardless of the economic theory employed. A monetarist no less than a Keynesian interpretation of economic fluctuations can lead to a fine-tuning approach to economic policy.

The persuasiveness of Keynes's view was greatly enhanced in Britain by historical experience, as well as by the example Keynes himself set. Britain retains an aristocratic structure—one in which noblesse oblige was more than a meaningless catchword. What has changed are the criteria for admission to the aristocracy—if not to a complete meritocracy, at least some way in that direction. Moreover, Britain's nineteenth-century laissez-faire policy produced a largely incorruptible civil service, with limited scope for action, but with great powers of decision within those limits. It also produced a law-obedient citizenry that was responsive to the actions of the elected officials operating in turn under the influence of the civil service. The welfare state of the twentieth

century has almost completely eroded both elements of this heritage. But that was not true when Keynes was forming his views, and during most of his public activity.

Keynes's own experience was also influential, particularly to economists. He set an example of a brilliant scholar who participated actively and effectively in the formulation of public policy—both through influencing public opinion and as a technical expert called on by the government for advice. He set an example also of a public-spirited and largely disinterested participant in the political process. And it is not irrelevant that he gained worldwide fame, and a private fortune, in the process.

The situation was very different in the United States. The United States is a democratic not an aristocratic society, as Tocqueville pointed out long ago. It has no tradition of an incorruptible or able civil service. Quite the contrary. The spoils system formed public attitudes far more than a supposedly non-political civil service. And it did so even after it had become very much emasculated in practice. As a result, Keynes's political bequest has been less effective in the United States than in Britain, which partly explains, I believe, why the "public choice" revolution in the analysis of politics occurred in the United States. Yet even in the United States, Keynes's political bequest has been tremendously effective. Certainly most writing by economists on public policy—as opposed to scientific and technical economics—has been consistent with it. Economists, myself included, have sought to discover how to manipulate the levers of power more effectively, and to persuade—or educate—governmental officials regarded as seeking to serve the public interest.

I conclude that Keynes's political bequest has done far more harm than his economic bequest and this for two reasons. First, whatever the economic analysis, benevolent dictatorship is likely sooner or later to lead to a totalitarian society. Second, Keynes's economic theories appealed to a group far broader than economists primarily because of their link to his political approach. Here again, Keynes, in his letter to Hayek, said it better than I can: "Moderate planning will be safe if those carrying it out are rightly orientated in their own minds and hearts to the moral issue. This is in fact already true of some of them. But the curse is that there is also an important section who could almost be said to want planning not in order to enjoy its fruits but because morally they hold ideas exactly the opposite of yours [i.e., Hayek's], and wish to serve not God but the devil. Reading the *New Statesman and Nation* one sometimes feels that those who write there, while they cannot safely oppose moderate planning, are really hoping in their hearts that it will not succeed; and so prejudice more violent action. They fear that if moderate measures are sufficiently successful, this will allow a reaction in what you think the right and they think the wrong moral direction. Perhaps I do them an injustice; but perhaps I do not."

Keynes did not let this analysis prevent him from serving until his death as chairman of the *New Statesman and Nation*—presumably in the hope of

influencing the moral views of its editors and writers. I regard Keynes's analysis as indicating that the key problem is not how to achieve a moral regeneration but rather how either to frustrate what Keynes regards as "bad morals," or to construct a political framework in which those "bad morals" serve not only the private but also the public interest, just as, in the economic market, private greed is converted to public service.

The literature on Keynes and on the *General Theory* is by now immense. Of the books specifically devoted to Keynes's life, two stand out: the initial authorized biography by his student and disciple, Roy F. Harrod, *The Life of John Maynard Keynes* (1951); and the more recent multi-volume biography by Robert J. A. Skidelsky, *John Maynard Keynes*, Vol. 1: *Hopes Betrayed, 1883–1920* (London: Macmillan, 1983), and Vol. 2: *The Economist as Prince, 1920–1937* (London: Macmillan, 1988). The *Collected Writings of John Maynard Keynes* have been published under the auspices of the Royal Economic Society in 29 volumes (Macmillan, 1971 to 1982), with a final *Bibliography and Index* yet to come. This splendid collection includes not only his major work but also his published articles on economics and politics, many previously unpublished items, including letters, official memoranda and notes, and the like.

The Organization of Private Payment Networks

John A. Weinberg

One of the key roles banks have traditionally played is in the execution of payments among participants in the economy. Liabilities issued by banks, such as demand deposits, are a primary means of payment. The widespread acceptability of such private liabilities requires a reliable method for the settlement of such obligations. In a world where people and economic activity are dispersed in space and time, settlement often requires a method for communication between locations where purchases of goods take place and those where payment liabilities are issued. That is, the use of bank liabilities as means of payment requires the support of an interbank network for clearing and settlement.

Throughout U.S. banking history, such multibank networks have played important roles. In New England during the Free Banking period (1836–1863), the system that was centered around the Suffolk Bank in Boston widened the area over which many banks' notes could circulate at par.[1] In the latter part of the nineteenth century, banks participated in clearinghouses for the clearing and settlement of local checks and had correspondent relationships to handle checks over greater distances.[2] While the Federal Reserve (the Fed) ultimately took over a large part of the clearing and settlement of checks, the banking industry has developed other private, multilateral networks for handling interbank payments. Most notable, perhaps, are the nationwide credit card associations.

Recently, interbank networks have received considerable attention. The ongoing growth and consolidation of Automated Teller Machine (ATM) networks has stimulated discussions in the academic and public policy communities of possible antitrust issues raised by such large joint ventures of banking organizations. Much of the discussion of possible public policy concerns regarding

This article has benefited greatly from the comments of Tom Humphrey, Jeff Lacker, Pierre-Daniel Sarte, and John Walter. The views expressed herein are the author's and do not represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] See, for instance, Calomiris and Kahn (1995).

[2] A recent description of check clearing in the nineteenth century is found in Gilbert and Summers (1996).

coming forms of electronic money centers on the network characteristics of these instruments.

This article takes the position that networks are fundamental to the role played by intermediary institutions in the payment system. Clearing and settlement, as the means of managing the financial relationship among individuals or institutions across time and distance, are inherently network services. The characteristics of network services have important consequences for the industrial organization of the payment system.

Arrangements for clearing and settlement of payments, whether private or public, involve an effect that is sometimes referred to as a network externality. Broadly, the private value to an individual of belonging to a network increases with the number of endpoints to which that network connects. Put differently, the private value an individual derives from participating in a network is that the individual can communicate with the other network members. At the same time, the individual's participation creates value for other members by adding to the number of endpoints.

There is an important difference between network externalities and other forms of externality. Perhaps the most common textbook externality is pollution; the economic activity of some individuals may produce pollutants that affect a much broader set of people. While individuals make choices about participation in the activity that generates pollution, they may have little choice as to whether to be affected by pollution. In the extreme case of the greenhouse effect, for instance, it may be impossible to avoid incurring the costs of pollution. Pollution is an external cost that has effects beyond the set of people engaged in the polluting activity. Network externalities, on the other hand, are more self-contained. An individual's decision to subscribe to a network creates external benefits for other subscribers by increasing the size of the network. Notice, however, that in order to enjoy the effects of the externality, one must join the network. The benefits to network participation, although partially external to the individual participant, are entirely internal to the network as a group. Since the group is composed of individuals engaged in mutually voluntary exchange with one another, one would expect the group's organization and pricing arrangements to take account of the "external" benefits associated with an individual's participation.

In a payment network, the value of communication between two endpoints is determined by the pattern of commerce. People at location A will place a high value on being in a payment network with people from location B if there is a high volume of commerce between the two locations. Since locations can vary widely in the sets of places their people go to shop, there can be variety among endpoints in both the private value of network participation and the external value that an endpoint creates for others through its participation. For a network to be sustainable, then, its services must be priced so that all of its intended members have an incentive to join. For a network to be efficient, it

must include all endpoints for which the total value of participation (private plus external value) exceeds the resource cost of participation, and only those endpoints.

This article proposes that the two standards of sustainability (which is defined more precisely below) and efficiency can form the basis of a positive theory of private, multilateral clearing and settlement arrangements.[3] Such a theory is quite distinct from the conventional view that network effects, as a form of externality, are a common source of market failure. This conventional view arises from the analysis of the behavior of network participants under the assumption that key organizational features of networks are exogenously fixed. By contrast, the theory presented below treats organizational arrangements as the endogenous outcomes of interactions among participants. Understanding the differences between these two theoretical perspectives is essential for understanding the role of central banks or other public entities in such activities.

The next section presents an abstract model of a network, gives some possible payment system interpretations, and shows how the essential network characteristics provide implications for network organizations. The following sections discuss some of the elements of a more general (and complete) model and apply some of the insights of the analysis to some historical and contemporary payment network issues. In particular, Section 3 discusses check clearing prior to the founding of the Federal Reserve and current issues involving ATM networks. In the former case, many observers have argued that check clearing was inefficient, as evidenced by the fact that checks sometimes followed very indirect routes in proceeding from initial deposit to final clearing. In the latter case, the use of surcharges (charged by an ATM-owning institution to depositors from other institutions) has been cited as an inefficient exploitation of market power in private networks. Section 3 will present the argument that the empirical facts of both these cases are consistent with a theory that predicts efficient private network organizations.

1. AN ECONOMIC MODEL OF A NETWORK

While network effects have often been said to be present in a variety of industries that are not explicitly networks, the focus here is on explicit networks.[4]

[3] Sharkey (1985) and Henriet and Moulin (1996), for instance, follow this approach to the theory of network structure and pricing.

[4] For instance, some authors have suggested the presence of such an externality in the market for personal computer operating systems; from a given set of alternatives, buyers prefer to have the system that is more widely chosen. In this case, the externality comes from the indirect effect that a system's popularity has on the likely availability of application software. Indeed, one might argue that a network effect is present in the retail grocery industry; the value to consumers of shopping at a larger store (or chain) might be enhanced by the store's ability to attract a wider set of suppliers, thereby offering the shopper greater variety. For a general survey of network effects, see Economides (1996).

In particular, the focus is on two-way networks where the underlying service for which people have a demand involves transmissions between two particular individuals. One can think of such a transmission as communication and the underlying product as a message sent from one specific individual to another. The model created below specifies the communication and consumption opportunities available to people in the economy.

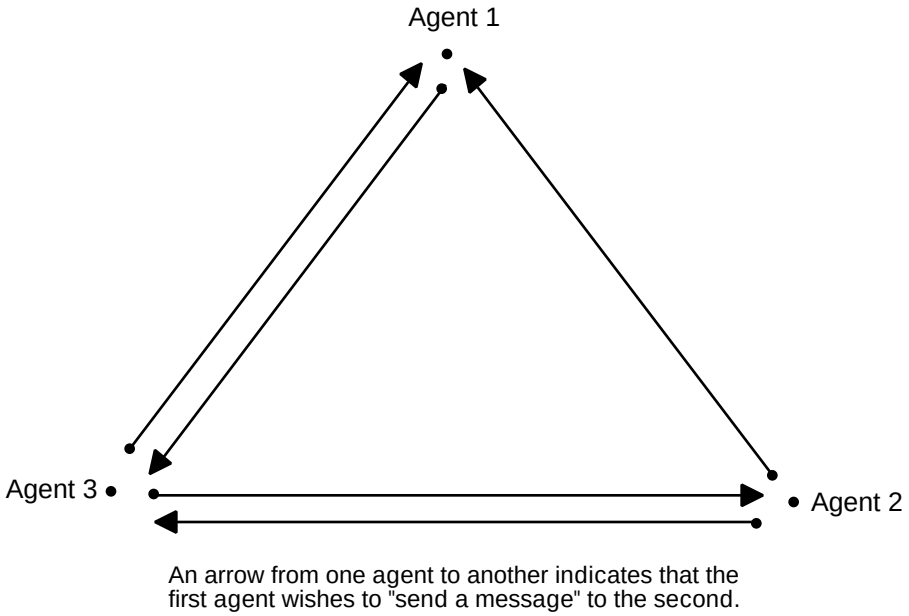
An Economy with Consumption and Communication

To create an economic model of a network, imagine a set of individuals, each of whom lives at a distinct location, separated from the others. One can imagine any number of individuals, and in general one might denote the set of individual agents as $\{1, 2, \ldots, N\}$. The key insights from the analysis can be gained from a simple example with three agents. Each individual derives utility from sending messages to some subset of the other individuals. Again, there is a wide array of possible patterns of desired communication. In particular, there may be heterogeneity in the overall value agents place on communication. Some agents may desire to send messages to a large number of recipients, while others may wish to communicate with only a few. The pattern of desired messages that is represented in Figure 1 has this feature. Agent 1 desires only to send a message to agent 3. Agent 2 derives utility from sending messages to agents 1 and 3, while agent 3 would like to send to agents 1 and 2. It is useful to assume that, in addition to communication, agents receive utility from consumption of a generic good and that each agent begins with an endowment of this good.

An agent's preferences for consumption and communication can be stated more formally as follows. Let J_i be the set of other agents with whom agent i wishes to communicate. Hence, in Figure 1, $J_1 = \{3\}$, $J_2 = \{1, 3\}$, and $J_3 = \{1, 2\}$. For any set J, use the notation $n(J)$ to denote the number of elements in that set; for instance, $n(J_3) = 2$. The agent receives utility v for each unit of communication, and utility is linear in consumption. If x_i denotes agent i's consumption of the generic good and U_i denotes the agent's utility, then $U_i = n(J_i)v + x_i$.

A model must also specify the technology available for communication. In particular, suppose there are two ways to communicate. A message can be sent by direct, bilateral communication at a cost (to the sender) of c_0, in units of the generic good. Alternatively, agents can buy access to a network. A network is a set of "connected" agents. If an agent is connected to a network, then messages to any other agent connected to the same network are costless. The cost, again in units of the generic good, of connecting an individual to a network is $c_s > c_0$. A more general specification of network costs might include a fixed (infrastructure) cost, a cost that is variable in the number of agents connected (like c_s per connected agent in the present specification) and a cost per communication (assumed zero here). The essential feature that is

Figure 1 A Simple Example of Demand for Network Services



captured by the simple specification given here is that, by expending resources to connect to one another, agents can reduce their costs per unit of communication. If we assume that the (utility) value of sending a message (v) is at least c_0, and that the agents' endowment of the consumption good is at least $2c_0$, then each agent will send all desired messages even in the absence of a network connection (the most any agent would spend on communication, if all communication was bilateral, is $2c_0$).

An efficient network is one which includes all agents whose private and external benefits of membership exceed the cost of connection and includes no other agents. To state this definition formally, one needs some additional notation. Let S be a possible network. That is, S is some subset of the agents. This set might also be termed a coalition. The private value to individual i of being a member of the connected set S depends on the number of other agents in S with whom i wishes to communicate. Letting J_i^S denote this set, the relevant number of connected agents for i is given by $n(J_i^S)$. Formally, J_i^S is the intersection of the sets S and J_i. Network membership does not change the actual amount of communication in which i engages; it is still worthwhile to communicate bilaterally with all who are in J_i but not in S. Therefore, the gross private benefit of membership (gross of connection costs) is the savings

in bilateral communication costs. For agent i, this value is $n(J_i^S)c_0$, since agent i wishes to send a single message to each of the other agents with whom he communicates.

In addition to the private value, agent i's membership in S creates value for other members of S. Define H_i as the set of agents who wish to communicate with i. Hence, in Figure 1, $H_1 = \{2,3\}$, $H_2 = \{3\}$, and $H_3 = \{1,2\}$. Agent i's membership in S creates value for all agents in H_i who are also in S. These agents are denoted H_i^S, the intersection of the sets H_i and S. The external value created by i's membership in S is $n(H_i^S)c_0$. With the private and external values of membership specified, one can state the following definition.

Definition 1: An efficient network is a set of agents S^* such that[5]

(1) $[n(J_i^S) + n(H_i^S)]c_0 \geq c_s$ for all i in S^*, and

(2) $[n(J_i^S) + n(H_i^S)]c_0 < c_s$ for all i not in S^*.

Condition (1) states that if agent i is in the network S^*, then the private-plus-external benefits from i's membership equal or exceed the connection cost. Condition (2) states that if private-plus-external benefits are less than the connection cost, then agent i is not in the network. Under the additional assumption that $3c_s < 5c_0$ (costs of universal connection are less than total communication costs in the absence of a network), an efficient network for the example of Figure 1 is one that includes all three agents.

Behavior of Agents in the Economy with an Exogenous Organizational Structure

Given an economic environment such as the one just described, how does one predict an outcome? In particular, what sort of network will emerge? Will it be efficient? How will agents share the costs and benefits of network connection? One approach to these questions is to assume a certain form of competition among potential networks. That is, one might assume a particular game played by the participants in the economy. To this end, it is useful to assume that there is an additional set of agents who have access to the network technology and derive utility only from the consumption good. These agents compete by offering network services to the agents who have a demand for communication. For simplicity, suppose that the services of these potential network providers are incompatible; each provider's network is unique and cannot be connected to other providers' networks. The "rules of the game" dictate the types of offers

[5] To be precise, the notation in Definition 1 should specify the particular network S^*. For instance, J_i^S should be $J_i^{S^*}$. When there is only a single network, as in Definition 1, suppressing the extra notation introduces no ambiguity. Further, for this economic environment the assumption of a single network is without loss of generality.

the sellers are allowed to make as well as the allowed responses by network users.

One possible game through which network providers could compete proceeds as follows. First, each potential provider announces an access price and stands ready to provide access to all comers at that price. Next, agents choose whether and from whom to buy access. Finally, communication and consumption take place. This game will be referred to as the *uniform price-setting game*. This name reflects the fact that sellers are not permitted to price discriminate by offering different prices to different buyers. The predicted outcome of a game is its Nash equilibrium, which is a set of strategies (price offers by providers and network choices by buyers) such that each player's (agent's) strategy is (privately) optimal, given the strategies of all the other players. Sellers in this game essentially bid for the right to provide network services to all who sign up. Since each potential provider is just as capable as all the others, price competition will tend to drive profits to zero. With uniform price setting, this competition can lead to extreme results, as in the following.

> Result 1: In the equilibrium of the uniform price-setting game for the environment described by Figure 1, no agent purchases network access, so all communication is bilateral.

To see that this result is true, note first that the connection fee must be at least c_s, since sellers have no incentive to sell at a loss. Agent 1, however, will not pay c_s to join the network, since joining saves him at most bilateral costs of c_0 (if agent 3 is in the network). Without agent 1 in the network, connection is not worth c_s to the other agents, so the equilibrium network is empty.

This result is a stark example of the general finding that under certain forms of competition equilibrium involves networks of inefficient size.[6] For the uniform price-setting game, such a result arises whenever there is at least one agent for whom the greatest possible private benefits of connection are less than the connection cost but for whom the private-plus-external benefits exceed costs. Such an agent is one who has a relatively low personal demand for communication but with whom many others wish to communicate (an agent for whom $n(J_i)$ is small and $n(H_i)$ is large).

An Alternative Model of Behavior: Sustainability

Results like Result 1 are useful for understanding the nature of competition in network markets. Their usefulness is limited, however, because they typically apply to specific games. The variety of games that might be played is virtually endless, even for the very simple economy of Figure 1. Would other games yield different results? In particular, are there games for which the equilibrium

[6] See Economides (1996) for a survey of such results.

network is efficient? The process of searching over all possible games is an impractical approach to the economic analysis of such environments. An alternative is to look directly at possible outcomes (allocations of network services and the consumption good) and ask which outcomes are sustainable, in a well-defined sense. In particular, suppose a certain allocation has been tentatively agreed upon by the agents in the economy. That allocation is sustainable if no subset of the agents can make themselves better off by allocating their own endowed resources among themselves. In the present environment, such a deviation from a proposed allocation would involve the subset of agents, or coalition, forming its own network and communicating with all other agents bilaterally; this coalition might also choose to reallocate its members' endowments of consumption goods, net of connection costs.[7]

To state this sustainability property more formally, an allocation in this economy can be defined as a network S and a payment for each agent, p_i, toward covering connection costs.[8] In considering possible deviations from proposed allocations by coalitions, it is sufficient to consider a coalition of any size that deviates to form a single network; if there were an incentive for a coalition to form two distinct networks, then each network would have its own incentive to form, regardless of whether the other forms. Letting $p = (p_1, p_2, \ldots, p_N)$, we can now define sustainability as follows.

Definition 2: An allocation (S, p) is sustainable if there does not exist a network S' and a payment allocation p' such that

(1) $p'_i + n(J_i - S')c_0 \leq p_i + n(J_i - J_i^S)c_0$ for each i in S', with strict inequality for some i in S', and

(2) $\sum_{i \text{ in } S'} p'_i \geq n(S')c_s$.

In the above definition, the notation $A - B$ for two sets A and B means all elements of A that are not in B (the complement of B on A). The first condition states that each agent in the deviating network S' is made better off by joining that network (and at least one is made strictly better off). An agent is made better off if his total outlays for bilateral communication and connection payments are reduced. The second condition is that the network collect enough in payments to pay for connecting all of its members.

One important fact to note about sustainability is that a sustainable allocation must be efficient. If it were not, a coalition consisting of an efficient network could form and arrange an allocation satisfying the two conditions above. In the case represented by Figure 1, a sustainable allocation will involve

[7] A sustainable allocation is defined here as an allocation in the *core* of the economy.

[8] In principle, one can allow for the existence of more than one network, although in this environment any allocation with more than one network is either equivalent to or (Pareto) dominated by an allocation with a single network.

a single network consisting of all three agents. The remaining task is to determine how the costs of connecting the agents ($3c_s$) should be shared among the three. As we have already seen, charging each agent c_s does not work; agent 1 can be charged no more than c_0. The remaining cost ($3c_s - c_0$) must be covered by charges to agents 2 and 3, with the restriction that neither can be charged more than $2c_0$, the cost to each of sending all messages bilaterally. Hence, one possible cost allocation is for agent 1 to pay c_0, agent 2 to pay $2c_0$, and agent 3 to pay $3(c_s - c_0)$. These prices give no agent an incentive to leave the network and send all communications bilaterally. The prices also leave no incentive for any pair of agents to form a separate, two-agent network and communicate with the third bilaterally. Hence, with these prices the efficient network is sustainable. This result is summarized below.

Result 2: In the case of Figure 1, a sustainable allocation has an efficient network $S = \{1, 2, 3\}$ and any payments (p_1, p_2, p_3) satisfying $p_1 \leq c_0$, $p_2 \leq 2c_0$, $p_3 \leq 2c_0$, and $p_1 + p_2 + p_3 = 3c_s$.

A Sustainable Allocation as the Outcome of a Game

Returning now to the question of competitive games played by potential providers of network services, does there exist a game under which the equilibrium network is the efficient network? Suppose, as before, that a large number of (incompatible) network service providers compete for subscribers. Instead of requiring that competition be only in the form of nondiscriminating access prices, suppose that the sellers can make any type of price offer they wish. That is, price offers can be in the form of a distinct price for each buyer. Refer to this game as the *perfectly discriminating price-setting game*. Equilibrium prices for this form of competitive bidding correspond to the sustainable cost allocations specified above.[9]

Result 3: Equilibria of the perfectly discriminating price-setting game are sustainable allocations.

To see this, suppose a seller has offered a set of prices satisfying the conditions stated in Result 2. Can any other seller offer prices that win customers from the first and yield a profit? In order to attract any individual buyer, a competing offer must give that buyer a lower price. In order to cover costs, however, the payment from at least one other buyer must be raised, so it is impossible to attract more than one buyer. If all buyers do not join, network connection is not attractive. Hence, a payment allocation satisfying the sustainability conditions cannot be undercut. On the other hand, any set of prices that

[9] There are many equilibria corresponding to many core allocations. In all of them, agent 1's access price is no greater than c_0, agents 2 and 3 face prices no greater than $2c_0$, and all agents connect to the network.

leaves a seller with strictly positive profits can be undercut. Accordingly, the sustainable allocations correspond with the set of equilibria.

The most notable feature of sustainable pricing arrangements is that, in order to support an efficient network, they require the subsidization of agent 1's connection by the other two agents; agent 1's connection fee must be less than the resource cost of connecting him. The other agents are willing to cover the remainder of the cost, because agent 1's (social) value to the network exceeds the (private) value he places on network access. This example illustrates a general point about arrangements that support efficient networks in environments in which agents are heterogeneous in the way they value network participation. Benefits (and possibly the costs) of network participation have a collective component. The key to sustaining an efficient network is in the distribution of these collective benefits and costs. This distribution must respect the capability of agents to leave the network, either individually or in groups. In a setting with heterogeneous agents, it can quite easily arise that the appropriate distribution of costs and benefits require that different agents pay different prices for essentially the same service.

The pricing arrangements that satisfy the conditions in Result 2 involve perfect price discrimination; they require that prices be tailored for each individual buyer of network services. Perfect discrimination is not always feasible. If, for instance, there is uncertainty about demands for network services and an individual's true demand characteristics are private information, then prices cannot be as finely targeted as in the above example. In this case, private information imposes further constraints on attainable allocations. It is possible to incorporate such constraints into the notion of sustainable arrangements. In such settings, pricing arrangements are likely to involve a less perfect form of price discrimination. For instance, prices for network services may be tied to observable characteristics or actions that are correlated with true demand. In an environment similar to the one discussed in this section, access prices that vary with the amount of communication might be able to achieve the desired price discrimination in a way that allows privately informed buyers to self-select among alternative pricing options.

2. ELEMENTS OF A GENERAL PAYMENT NETWORK MODEL

The model and example of the previous section were specified in terms of a generic communication service. The same sort of network structure, however, can arise in a model that is specified in such a way as to capture important aspects of payment system markets. Any noncash payment mechanism is a communication network in a fundamental sense. An instrument presented in payment for goods or services is an instruction to transfer monetary value from the buyer's to the seller's ownership. Execution of such an instruction requires

communication between the point of sale and the location or institution at which the buyer's value is held. This section sketches some of the ingredients of a general payment network model. The key point is that the private and external values to individuals of being connected to a payment network depend on the underlying pattern of commerce.

An Economy with Payment Services

As in the above section, suppose that there is a set of N distinct locations at which agents live and economic activity takes place. Unlike the previous section, suppose that there is a large number of agents living at each location. Agents consume two types of goods: a generic good and location-specific goods. Different people have different preferences for location-specific goods. In particular, each agent desires the specific good from exactly one location. One might, for instance, denote by ϕ_{ij} the fraction of agents from location i who wish to consume the specialized good at location j. These fractions determine the economy's pattern of commerce. Agents travel from their home location to the locations at which they wish to consume and purchase location-specific goods with claims on amounts of generic goods (or with the generic good itself). In some environments, one might also imagine that these transactions are made using government-issued fiat currency.

Making transactions across locations is costly. This cost might arise from a number of frictions. If debt claims are created in the purchase of location-specific goods, then there may be costs associated with communicating information about these claims across locations or in making final payment. If buyers carry the generic good with them to make purchases, there may be transportation costs or other losses incurred on the way. Finally, if traveling for consumption takes time and buyers carry non-interest-bearing currency for transaction purposes, then there may be a seigniorage cost associated with location-specific consumption. The specific nature of the costs depends on the details of the economic environment.

As above, suppose that there is a technology for connecting locations in a network for the purpose of clearing and settling payment obligations. While connection may involve a fixed cost, the variable cost of making transactions among connected locations is lower than between locations that are not connected. For instance, if payment for location-specific goods requires shipment of generic goods, a network that allows multilateral communication and calculation of net obligations may economize on shipping costs. Alternatively, in a monetary economy, a network that allows people to substitute debt claims for currency may allow agents to save on seigniorage costs.

The value to agents at a particular location of being connected to the network depends on which other locations are connected. In particular, agents at location i place a high value on being in a network that includes locations j for which the fractions ϕ_{ij} are large. By the same token, a location at which

many people want to shop (j such that ϕ_{ij} is large for many i) is one that brings a high external value to any network it joins. Hence the notion of sustainable pricing of network services, as presented in the previous section, will imply that these popular locations receive preferential treatment in the pricing of network services. As is the case for location 1 in the example of Figure 1, there could easily be locations for which the private value of network services is small while the external value is large. These would be locations that attract many consumers from elsewhere but whose residents consume mostly their own location-specific goods. Pricing to support an efficient network could require that such locations pay less than the cost of connecting them, as is the case in the example.

Payments-related models that have the features outlined in this section will have the same implications as the example from the previous section. Network industries will tend to be organized in ways that achieve sustainable allocations. This means, first, that there is a tendency toward efficient network structures. Second, the sharing of the costs and benefits of network organizations must respect the ability of participants to form or join alternative organizations. Accordingly, network members who bring large external benefits to other members may need to receive a share of the net benefits that appears to be out of line with those members' own use of the network services. Such an impression mistakenly focuses only on an individual's private benefits of network participation and not the benefits that an individual brings to other participants.

Barriers to Competition

An important maintained assumption in the foregoing discussion is that any agent or group of agents that is dissatisfied with an arrangement is free to pursue an alternative. That is, there are no barriers to entry. Various types of barriers might arise in economic environments. For instance, if all agents do not have access to the same technological capabilities, then it might be difficult for agents that are dissatisfied with their network services to set up or seek out an alternative network. Also, there may be investments in network provision or participation that, once made, represent sunk costs. A sunk cost is a cost that cannot be fully recovered. In this case, an incumbent network would have a cost advantage over a competitor; while the competitor must incur the sunk costs, those costs are no longer part of the incumbent's decision calculus.

Other barriers to competition might arise from legal restrictions. For instance, if sellers were to face a legal prohibition of price discrimination, then the types of network services arrangements they could offer would be sharply limited; as seen above, price discrimination can be essential for the efficient provision of network services with heterogeneous buyers. Other legal barriers might take the form of restrictions on which particular sellers can offer which particular services.

A final form of potential barrier that is worth noting arises from the behavior of sellers themselves. An incumbent seller of network services might attempt to impose rules on its buyers that make it difficult for them to switch to a competing service. The possibility of restrictive rules set by network providers has been a subject of interest in recent policy discussions concerning ATM network mergers.

If there were barriers to competition, then an inefficient network structure could persist. In such a case, can public policy intervention improve on private market performance? The answer depends on the source of the barriers. In the United States and other developed economies, enforcement of antitrust laws is in part intended to guard against the anticompetitive use of restrictive rules by sellers in their contracts with buyers. In cases where a barrier to competition is the result of a legal restriction on the behavior of market participants, such restrictions might have other public purposes. Here, as in the case where barriers may have technological sources, it may be difficult or impossible for public intervention to remove the barriers. In these cases, an incumbent provider might extract monopoly rents, for instance, by inefficiently limiting network size.[10] In such cases regulation of the pricing and product offerings of an incumbent seller might be useful in promoting network efficiency. Governments in many economies have traditionally taken this approach to telecommunications markets.

3. TWO APPLICATIONS

One can apply the logic presented above to a number of actual payment network examples. This section will provide a brief discussion of two such examples, one historical and one current. The historical example involves the process of clearing checks in the United States in the period prior to the founding of the Federal Reserve System. The current example involves the growing geographical reach of multibank ATM networks.

Check Clearing before the Federal Reserve

By the late nineteenth century, checks had already become a dominant form of payment. As the banking industry was highly fragmented, a significant portion of all checks were interbank checks. Clearing of checks between the depositor's and the check writer's bank occurred in one of a number of ways.[11] A bank

[10] Of course, even a monopoly immune to competition will not necessarily produce inefficient results. If the monopolist has sufficient ability to price discriminate, monopoly behavior can approach full efficiency. In this case, the monopolist's rents come at the expense of consumer welfare but not at the expense of total welfare.

[11] For a recent discussion of pre-Fed check clearing see Gilbert and Summers (1996). A classic detailed account is found in Spahr (1926).

holding checks drawn on accounts at another bank could present the items directly, in person, at the paying bank. By law, the paying bank was required to make payment on such checks without imposing any presentment fee. On the other hand, checks that were presented through the mail could be subject to a fee imposed by the paying bank. Banks could also clear checks through the services of an intermediary, or correspondent bank. Banks with a correspondent relationship might have entered into a mutual agreement to accept checks from one another without imposing fees.

At a time when bank branching was limited by law, correspondent relationships were particularly important for clearing checks in cases where the paying bank was relatively distant from the bank in which the check was initially deposited. If both banks had correspondent relationships with the same intermediary bank, then collection of the item could proceed free of fees. In this sense, any two banks that were connected by a chain of correspondent relationships belonged to a network. How was the composition of such networks likely to have been determined? The value to a bank of belonging to a network depended on the frequency with which the bank received checks drawn on other members of the network. Further, if there were particular institutions that had relatively frequent and large volume interactions with many other banks, then such institutions would naturally serve as central intermediaries in a correspondent network. For instance, a city bank might deal with a number of country banks in the surrounding region. The city bank might, in turn, maintain relationships with banks in other cities that serve as correspondents for their regions. This organization of a check-clearing network could economize on the costs of shipping checks. A small bank in a remote town could, for instance, send a single shipment of all its out-of-town checks to its correspondent, with the links between larger correspondents serving as "trunk lines."

In a correspondent network like that described above, consider the problem faced by a bank that receives a check drawn on an institution with which it rarely deals. The receiving bank could send the item directly to the paying bank. In this case, however, the paying bank might charge a fee for presentment. Alternatively, the receiving bank could send the item, along with its usual shipment, to its correspondent bank. Then, through a chain of correspondent relationships, the check might ultimately be paid at its par value (with no presentment fee). This indirect alternative has two potential sources of savings. First, presentment fees might be avoided. Second, there may be savings on the costs of transporting checks. The marginal cost of adding an item to a routine shipment is virtually zero, certainly smaller than the postage cost of sending a single item directly. Indeed, similar economies may have been available at the receiving end of check shipments. A paying bank may have found it more convenient and cost effective to receive and process bundles of checks sent by intermediaries with whom it had a standing relationship.

The Circuitous Routing of Checks

The history of check clearing during the period that preceded the founding of the Fed contains examples of checks traveling over circuitous routes to get from the banks in which they were initially deposited to the paying banks. A bank of first deposit, for example, might have sent a check to its correspondent located to the east even though the paying bank was located to the west. There are two possible interpretations of such examples. On the one hand, such instances might provide evidence of an inefficiency created by paying banks' ability to assess presentment fees. On the other hand, such cases could be consistent with the operation of an efficient correspondent network. The pattern of links in the network (correspondent relationships) was determined by the usual pattern of commerce. The occasional circuitous route for check clearing resulted simply because there were occasional exceptions to the usual pattern of commerce. Given the existing links, it was efficient to send these occasional items together with routine shipments. Indeed, in this view, it is possible that presentment fees reinforced network efficiency by reducing the incentive for individual banks to bypass the network.

The second interpretation is consistent with the analytical framework suggested above. Under this interpretation, presentment fees may have actually reinforced efficient check-clearing relationships by discouraging the direct presentment of occasional, solitary items. Those few items that were sent directly and on which presentment fees were paid are likely to have been items for which the bank of first deposit could not foresee a sufficiently reliable chain of correspondent relationships. That is, these were items for which the bank that was due payment had little alternative to direct presentment (through the mail). Accordingly, such items would constitute a market segment (in the market for clearing by direct presentment through the mail) with relatively inelastic demand. The efficiency cost of charging a high price (above marginal cost) to market segments with inelastic demand is relatively small; with inelastic demand, quantity purchased does not decline much as the price is raised. In other words, presentment fees allowed paying banks to price discriminate between institutions with good alternatives to direct presentment and those without such alternatives. It is entirely possible that the effects of such discrimination were primarily distributional. While some buyers gain at the expense of others, price discrimination typically increases the exploitation of gains from trade relative to uniform pricing in the presence of market power.[12]

ATM Networks

An ATM transaction, like a check transaction, can require interbank clearing and settlement. When the holder of an ATM card issued by one bank makes a

[12] Each paying bank had some market power in the sense that it was the only bank that could provide final settlement of a check.

withdrawal at an ATM owned by another bank (or perhaps by a nonbank business), the transaction must be cleared by communicating information about the withdrawal to the card-issuing bank and settled with an appropriate transfer of funds from the issuer to the ATM owner. In recent years, multibank ATM networks have become increasingly important in allowing cardholders to access their funds at ever widening sets of locations.[13] A regional ATM network is usually organized as a joint venture owned by some or all of its participating banks. The network typically has a brand name that is placed on its members' machines and cards.

Membership in an ATM network is valuable to a bank mainly because access to the network's set of ATM locations enhances the value of the ATM services the bank offers its depositors. Clearly, membership is more valuable the more extensive the network's set of locations. A bank's membership also brings "external" benefits to other members by adding to the set of locations. Some locations, however, are more valuable than others. For instance, banks (and their customers) may place a high value on having access to an ATM at a vacation resort. Such a location may be one for which the external benefits of network participation are greater than the private benefits to the owner of the particular ATM. The theory presented above suggests that a sustainable network will need to price its services or share its profits so as to allow the ATM owner to realize some of the external benefits of its participation.

Surcharges

There are a number of ways in which network arrangements could induce participation of institutions that bring large external benefits. For instance, if the network imposes membership fees, lower fees could be set for members with more desirable locations. Similarly, if the network is organized as a joint venture, then arrangements for profit sharing could conceivably reflect differences in the values of members' locations. Since the values of locations are likely to be related to the intensity of ATM use, prices based on transactions might also be used for allocating network costs and benefits. In ATM networks, when a cardholder from one member uses the ATM of another member, the cardholder's bank pays an "interchange" fee to the ATM owner. Since this fee is typically set by the network and is usually a uniform, per-transaction charge, its flexibility for cost and benefit sharing is limited. Owners might also be able to realize some of the network benefits of desirable locations by imposing a transaction fee directly on the cardholder. Cardholders' willingness to pay such fees, known as surcharges, would depend on the value of having access to cash at the particular location and on the degree of competition for cash access services at that location.

[13] A description and history of ATM networks are found in Baker (1996).

Many regional ATM networks, as well as the national networks owned by the Visa and Mastercard associations, once had implicit or explicit agreements among their members banning surcharges. These restrictions were ultimately challenged, often by owners of ATMs at high-value locations.[14] The debate over surcharges centers on two opposing interpretations of the role such fees play in the market. One interpretation holds that a ban on surcharges prevents ATM owners from abusing the monopoly power they gain from having desirable locations. The other interpretation is suggested by the arguments presented in this article; surcharges support the formation of efficient networks by allowing participants who bring large external benefits to the network to capture some of those benefits. Under this second view, a ban on surcharges is an attempt by network owners to impose a certain distribution of network benefits, a distribution that may not be sustainable.

Does the second view imply that monopoly rents earned by ATM owners with particularly advantageous locations do not create the inefficiency usually associated with monopoly power? Not necessarily. If a network includes some truly unique locations and there is no possibility of entry by a competitor at those locations, then the unique locations are essentially natural monopolies. Banning surcharges does not eliminate the rents from those monopoly positions. Rather, a ban is an attempt to spread those rents among all the banks in the network. If a network includes valuable locations offering equal access for customers of all network members but not for nonmembers, then members can extract rents in the fees they charge their customers for account services that include ATM access. In this case, the rents would be extracted from all customers, even those with no demand for access to the highly valued locations. Allowing surcharges, on the other hand, allows monopoly rents to be collected in a more discriminatory fashion. Such price discrimination can have the effect of reducing the inefficiency from monopoly power.

A second important point about monopoly rents is that an ATM owner's ability to extract rents is limited by competitors' ability to enter monopolized market segments. Hence, the primary public policy concern with market power should be "How is it maintained?" rather than "How is it exploited?" Competitors' incentives to enter markets by placing ATMs at particular locations is greatest when owners can earn location-specific rents. Hence, a ban on surcharges creates a situation in which incentives to engage in location-specific competition are muted.

[14] See for instance *Bank Network News,* September 13, 1995.

4. CONCLUSION

The theoretical framework presented above suggests that the organization of networks is driven by the desire of market participants to devise sustainable multilateral arrangements. In both of the cases discussed in the previous section, the framework leads one to interpret observed network structures and pricing arrangements as components of an efficient arrangement. Such an arrangement must take into consideration both the private value that participants derive from the network and the external benefits that participants bring to the network. Hence, in a world with heterogeneous demands for network services, price discrimination and other means of "unevenly" distributing the net benefits of network services can be essential for supporting efficient network structures.

Does this article's framework necessarily imply that all actual networks are efficient and that public intervention can never improve upon the economic performance of a private network? As discussed in Section 3 above, this best-of-all-possible-worlds result holds strictly only when there are no barriers that prevent groups of economic agents from pursuing the alternatives of their choice. The role of public policy, therefore, is to understand the sources of barriers to such choice. There may be some cases in which barriers are technological and cannot be overcome. In these cases, some regulation of pricing practices might be called for. This argument, however, is nothing more than the traditional justification of regulation of a natural monopoly. In other cases, barriers may be created through the rules imposed by incumbent providers of network services in an attempt to preserve market share. Public intervention to eliminate such rules could be beneficial. This argument closely mirrors traditional justifications for antitrust scrutiny of the conduct of firms with large market shares. In all cases, the framework begins with a presumption of efficiency. This presumption is expressed by the question, "If all economic decisionmakers are always free to make alternative arrangements, why wouldn't the arrangement on which they actually agree be efficient?" This presumption seems also to be a good place for public policy to begin.

REFERENCES

Baker, Donald I. "Shared ATM Networks: the Antitrust Dimension," *The Antitrust Bulletin*, vol. 41 (Summer 1996), pp. 399–425.

Bank Network News. "New Math Renews Old Surcharge Debate," September 13, 1995, p. 1.

Calomiris, Charles W., and Charles M. Kahn. "The Efficiency of Self-Regulated Payment Systems: Learning from the Suffolk System," *Journal of Money, Credit, and Banking*, vol. 28 (November 1996), pp. 766–97.

Economides, Nicholas. "The Economics of Networks," *International Journal of Industrial Organization*, vol. 14 (October 1996), pp. 673–99.

Gilbert, R. Alton, and Bruce J. Summers. "Clearing and Settlement of U.S. Dollar Payments: Back to the Future?" Federal Reserve Bank of St. Louis *Review*, vol. 78 (September/October 1996), pp. 3–27.

Henriet, Dominique, and Herve Moulin. "Traffic-Based Cost Allocation in a Network," *RAND Journal of Economics*, vol. 27 (Summer 1996), pp. 332–45.

Sharkey, William W. "Economic and Game Theoretic Issues Associated with Cost Allocation in a Telecommunications Network," in H. Peyton Young, ed., *Cost Allocation: Methods, Principles and Applications*. New York: North-Holland, 1985.

Spahr, Walter. *The Clearing and Collection of Checks*. New York: The Bankers Publishing Co., 1926.

The Case for a Monetary Rule in a Constitutional Democracy

Robert L. Hetzel

Constitutional democracy protects individual liberty. It does so by placing restraints on the arbitrary exercise of power by government. A primary restraint is the constitutional protection of property rights. The monetary arrangements of a country either promote or undermine that protection.

Money is unique in that its value in exchange far exceeds the cost of producing an additional unit. On the one hand, governments have an incentive to print additional money to gain "free" resources, or seigniorage revenues.[1] On the other hand, the central bank must limit the quantity of money in circulation to control prices.

Through its influence on seigniorage, money creation affects how government raises revenue. It can also affect who within government decides how that revenue is spent. Through its influence on fluctuations in the price level, money creation influences the extent of arbitrary redistributions of wealth among individuals. The institutional arrangements that govern the creation of money then bear on two aspects of the protection of property rights: the taking and disposition of wealth from the public and the distribution of wealth by government between individuals.

The opinions expressed herein are the author's and do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Under a commodity standard, the Mint had a monopoly over coinage. It charged a fee called seigniorage for turning bullion into coins. Today, seigniorage is the general term for the revenues a government gains from its monopoly over the creation of fiat money. The resources commanded by a paper dollar (or its electronic equivalent bank reserves) far exceed the resources needed to create that dollar. The central bank varies the amount of revenue it raises for the government through seigniorage by varying the rate of money creation and, consequently, the rate of inflation. Up to a point, higher inflation yields more revenue. When a central bank allows high rates of money growth and inflation, seigniorage is commonly referred to as an "inflation tax." The holders of cash, who must add to their cash balances to restore the purchasing power eroded by inflation, pay this tax.

A legislative mandate from Congress requiring the Federal Reserve (the Fed) to stabilize the price level and to hold only government securities in its portfolio would complement the rules in a constitutional democracy that protect property rights.

A *Financial Times* (1990) editorial made the case for rules in the conduct of monetary policy:[2]

> The notion that money must fall within the domain of day-to-day politics is a 20th century heresy. . . . Painful experience with the modern manipulation of monetary policy suggests that money is more appropriately an element of the constitutional framework of democracy than an object of the political struggle. Monetary stability is a necessary condition for a working market economy, which is itself a basis for a stable democracy.

1. OVERVIEW

Sections 2 and 3 provide historical background showing that monetary arrangements in Britain and the United States have in the past raised issues basic to the design of a constitutional democracy. Early British experience influenced American thinking during the Revolutionary War, in which protection of property rights was a central issue. Section 3 reviews how during the period of the Articles of Confederation the states' discretionary control of money undermined property rights. The monetary abuses of this period contributed to the convening of the Constitutional Convention.

Section 4 reviews the more recent experience with discretionary control of inflation in the 1970s and argues that this experience produced the same arbitrary redistributions of wealth that characterized the earlier period. The earlier experience led the Founding Fathers to remove government discretion over money creation by putting the United States on a specie (gold or silver) standard. The later experience caused the Fed in practice to assign priority to maintaining a low rate of inflation. However, in this more recent period, the United States has not put in place institutional arrangements to ensure monetary stability in the future. Section 5 reviews fiscal transfers by the Fed made possible through seigniorage and discusses how they can circumvent the constitutionally mandated budget process. Section 6 offers concluding comments.

2. PRINCIPLES OF CONSTITUTIONAL DEMOCRACY

The following historical review provides examples of how monetary arrangements either reinforce or relax constitutional constraints on the exercise of power by government. Discretionary control over inflation can undermine the

[2] For an economic argument in favor of rules, see Friedman (1960) and Lucas (1983). For a constitutional argument, see Friedman (1962).

accountability provided for in a constitutional democracy. The power of a government to increase the revenue from seigniorage through inflation without explicit legislation lessens democratic accountability. Moreover, the ability to allocate seigniorage revenues discretionarily, that is, outside the constitutionally mandated budget process, lessens accountability.

British Constitutional Democracy

The germ of constitutional democracy appeared in 1215 when British nobles forced King John to sign the Magna Carta. The Magna Carta established the principle that no one is above the law. Moreover, it articulated two principles that bear on monetary arrangements. First, government should be divided into separate parts capable of counterbalancing each other. Second, taxation "shall be levied in our kingdom only by the common counsel of our kingdom." The practical working out of these principles required decisions about which part of government would control and allocate seigniorage.

The Glorious Revolution in 1688 gave Britain its present parliamentary form of constitutional democracy. The competition between Crown and Parliament for political power, which Parliament won, was driven in part by religion and in part by Parliament's desire to gain secure property rights. The desire by Parliament to stop the arbitrary seizure of property by the Crown led to the particular characteristics of British constitutional democracy that inspired Americans in their Revolution. The exclusive right of Parliament to levy taxes was especially important.

In the past, some English kings had debased the coinage as a way of obtaining revenue. Despite protests by Parliament, Henry VIII had regularly lowered the precious metal content of coins. Debasement of the currency gave the Crown a source of income independent of Parliament. After the Glorious Revolution, Parliament took over the Mint to foreclose debasement as a source of revenue. When Parliament incorporated the Bank of England in 1694, it prohibited the Bank from lending to the Crown without Parliament's consent.

American Revolution

John Locke was the philosopher of the Glorious Revolution. He understood the two major incentives to the arbitrary exercise of power: the seizure of political power and the seizure of property. Locke ([1690] 1986, p. 180) wrote that men "join in society with others . . . for the mutual preservation of their lives, liberties, and . . . property."

Locke included popular consent to taxes as a natural right. That principle, expressed in the phrase "No Taxation without Representation," became the rallying cry of the American Revolution. The Declaration of Rights in the Virginia state constitution, written by George Mason and adopted in 1776, states that "men . . . cannot be taxed or deprived of their property for publick uses, without their own consent, or that of their representatives so elected" (Commager 1962,

p. 104). In the Declaration of Independence, Thomas Jefferson listed as one of the "repeated injuries and usurpations" of King George his assent to "imposing Taxes on us without our Consent."

The arbitrary seizure of property by the British Crown became a major factor precipitating the American Revolution. James Otis denounced writs of assistance, which gave the Crown's agents broad authority to search houses and confiscate smuggled goods not restricted by "probable ground" or particular "houses specially named" (Commager 1963, pp. 46–47). The Constitution provided for the protection of property rights in the Fourth Amendment, which guaranteed the right of "people to be secure in their persons, houses, papers, and effects against unreasonable searches and seizures," and in the Fifth Amendment, which provided that no person shall "be deprived of life, liberty, or property, without due process of law; nor shall private property be taken for public use, without just compensation." As explained below, the writers of the Constitution viewed monetary arrangements as part of the constitutional protection of property rights.

3. THE CONSTITUTIONAL PROTECTION OF PROPERTY RIGHTS

Abuse of Money Creation under the Articles of Confederation

In the period of the Confederation following the Revolutionary War, actual or threatened civil disorder in several states, especially Massachusetts, prompted calls for a convention to create an effective central government. The political difficulty of imposing the excise taxes necessary to pay debts inherited from the Revolutionary War caused many states to yield to pressure to repay debts through the issue of paper money. Rhode Island effected a general transfer of wealth from creditors to debtors through the inflationary issue of paper money combined with legal tender laws forcing creditors to accept paper money rather than specie. In a number of states, legislators used inflation in combination with price controls to transfer income from creditors, merchants, and large planters to farmers, debtors, and artisans. On August 7, 1786, James Madison (Rutland 1975, p. 89) wrote:

> . . . the States are running mad after paper money, which among other evils disables them from all contributions of specie for paying public debts, particularly the foreign one. In Rhode Island a large sum has been struck and made a tender, and a severe penalty imposed on any attempt to discriminate between it and coin. The consequence is that provisions are withheld from the Market, the Shops shut up—a general distress and tumultuous meetings.

In a letter to Jefferson, Madison complained about the "warfare & retaliation" that arose among states due to laws allowing citizens of one state to pay out-of-state debts in depreciated paper money (Rutland 1975, pp. 94–95). In

Federalist essay 10 (Beloff 1987, p. 47), Madison expressed his famous idea that in a large republic "a greater variety of parties [would militate] against the event of any one party being able to outnumber and oppress the rest." One reason was that in a large republic "a rage for paper money, for an abolition of debts, for an equal division of property, or for any other improper or wicked project, will be less apt to pervade the whole body of the union, than a particular member of it" (Beloff 1987, p. 47).

The 1786 disturbance in Massachusetts called Shay's Rebellion acted as a catalyst for the convening of a Constitutional Convention. Armed farmers closed courts to prevent foreclosures on properties for tax delinquency and threatened to march on Boston to force passage of easy money legislation. Such threats prompted the Constitutional Convention to protect property rights by prohibiting states from issuing paper money, from making legal tender laws, and from impairing the obligation of contracts. In *Federalist* essay 44 (Beloff 1987, p. 227), Madison defended the Constitution's prohibition of the issuance of paper money (then called bills of credit) by the states:

> The extension of the prohibition to bills of credit must give pleasure to every citizen in proportion to his love of justice and his knowledge of the true springs of public prosperity. The loss which America has sustained since the peace from the pestilent effects of paper money on the necessary confidence between man and man; on the necessary confidence in the public councils; on the industry and morals of the people, and on the character of republican government, constitutes an enormous debt against the states.

Jefferson believed that the Constitution denied to Congress the "power of making paper money or anything else a legal tender" (Lipscomb 1903, p. 65). He wrote that paper money "is liable to be abused, has been, is, and forever will be abused, in every country in which it is permitted" (Ford 1898, p. 416). The historian Jack Rakove (1996, p. 44) argues that Madison formulated his views on constitutional government in response to the arbitrary redistributions of wealth caused by the paper money inflations and legal tender laws of the states in the Articles of Confederation period.

Limiting the Power of Government to Seize Property Arbitrarily

The Founding Fathers carefully balanced the need for government to raise revenue with the need to protect private property from arbitrary seizure. They did so by entrusting fiscal policy to Congress, which by design encourages open debate.[3] That debate allows for the monitoring that gives substance to

[3] Congress is composed of a large number of members. There are two houses of Congress, each with its own subcommittees, committees, and rules committee. In order to become law, a bill must pass through all these individual centers of power. Along the way, a bill can be killed by a filibuster. A bill must also go through a conference committee to reconcile differences between each house of Congress and then survive a possible presidential veto. This convoluted structure not only disperses power, but also enhances public discussion. The debate and controversy

the principle that sovereignty resides with the people. Tax legislation must originate in the House, whose members are elected every two years. No funds can be spent by the other branches of government without explicit authorization by Congress. Article I, Section 9, of the Constitution states, "No money shall be drawn from the Treasury, but in consequence of appropriations made by law; and a regular statement and account of the receipts and expenditures of all public money shall be published from time to time." Article I, Section 8, enforces congressional control of fiscal policy by allowing only Congress "to borrow money on the credit of the United States." To enforce the separation of powers, however, a different branch of government, the Executive Branch, spends the revenues raised by Congress.

In Article I, Section 8, of the Constitution, the Founding Fathers assigned to Congress the responsibility "to coin money, regulate the value thereof." By assigning to Congress the responsibility to determine the metallic content of coins, they prevented the President from following the example of European kings and debasing the currency. Article I, Section 10, prohibited state governments from printing paper money. Taken together with Article I, Section 8, the Founding Fathers put the country on a specie standard. In this way, the authors of the Constitution sought to limit the ability of government to abrogate contracts by taking actions affecting the price level.

Monetary arrangements were part of the initial constitutional framework. By removing money creation and the determination of the price level from the political process, the Founders limited the ability of government to purposefully redistribute wealth through inflation and inflation combined with price controls. Those same restrictions also protected property rights by preventing recourse by the government to the unlegislated tax of inflation.

However, the specie standard ultimately did not survive as part of the constitutional framework. At times, the specie standard allowed for inflation or deflation and was itself a source of instability. A mandate from Congress to the Fed to stabilize the price level would possess the advantage of the specie standard of eliminating discretionary control over the price level. It would complement other protections afforded property rights. At the same time, such a mandate would avoid the instabilities of the specie standard.

The argument for a rule requiring the Fed to stabilize the price level is not one of original constitutional intent. And it is not a legal argument about the constitutionality of discretionary as opposed to rule-based monetary arrangements. The Constitution does not require government to stabilize the price level. Money is not a capital "C" constitutional issue involving a legal interpretation of the Constitution.

---

generated by the multistep procedure of requiring repeated majorities among groups with widely differing self-interests generates news and thus supplies the information citizens need to monitor government. It also reduces the possibility of precipitate action.

However, the Constitution does provide broadly for the protection of property rights. A mandate from Congress to the Fed requiring price-level stabilization would buttress that broad constitutional protection. Such a rule is desirable for the same reason that rules in general are desirable in a constitutional democracy: to constrain the arbitrary exercise of power by government. In this sense, the sense used here, money is a small "c" constitutional issue referring to a form of government where rules limit government power. (See the Appendix for a brief history of money and the Constitution.)

4. THE MODERN EXPERIMENT

Although monetary arrangements have disappeared as part of the constitutional order, Congress remains constitutionally responsible for the value of the dollar. While it has delegated that responsibility to the Fed, it has done so without clear instructions on the desirable behavior of the price level.

The Inflation of the 1970s

In the 1960s and 1970s, the combination of pressures to make the economy grow rapidly and a belief that inflation arose from factors unrelated to money growth caused the Fed to pursue an inflationary monetary policy. Inflation generated an unlegislated transfer of resources to the government through an increase in seigniorage. More important than seigniorage, inflation generated government revenue through its interaction with a tax code not indexed for inflation. The arbitrary redistributions of wealth due to the inflation of this period recall the earlier experience of the Articles of Confederation period.

Prior to indexing of the personal income tax brackets and exemptions in 1985, inflation combined with progressive tax rates to push individuals into higher tax brackets. With inflation, capital gains taxes still fall not only on the real gains from selling an asset, but also on the paper gains that compensate for inflation. In addition, inflation raises corporate tax rates by lowering the real value of historical depreciation costs and by raising the measured, but not real, profits on holdings of inventories. Inflation also erodes the value of the estate tax exemption. (See Feldstein [1996] and the Appendix in Hetzel [1990].) The interaction of inflation and a tax code specified in current dollar terms created enormous uncertainty and distortions in the 1970s. Corporations favored short-term over long-term investment. Investors shifted capital out of the corporate sector and into real estate in response to the effective rise in corporate tax rates.

Inflation allowed a shadow fiscal system where the combination of inflation and price controls transferred income to politically influential groups. Such transfers were not subject to the safeguards provided by the public discussion that accompanies explicit legislation to impose a tax. The housing industry successfully lobbied for the use of Regulation Q to limit the interest rates

that depository institutions could pay on savings and small-denomination time deposits. As a result, when inflation rose above Regulation Q ceilings, savers of modest means received negative returns. Wealthy investors were unaffected as they could invest directly in money market assets whose yields incorporated an inflation premium. By not allowing the usury ceilings on the interest rates paid on loans to rise with inflation, state governments discriminated against savers. Rent controls, especially in New York and California, appropriated the wealth of apartment owners. Price controls on domestically produced oil, instituted as part of the Nixon price controls in early 1974 but kept until the early 1980s, rewarded small, politically influential refiners.

Inflation and Scapegoating

When discussing the costs of inflation, economists concentrate on economic costs, such as the increased number of trips individuals make to the bank to economize on the holding of cash balances. The actual experience with the inflation of the 1960s and 1970s, however, demonstrated that the problems created by inflation can extend beyond the purely economic.

Inflation threatened the dispersion of power characteristic of the U.S. political system. It did so by creating a demand for immediate action to deal with a problem that could only be solved through a long period of perseverance and patience. Moreover, inflation created an incentive for the political system to identify scapegoats and to deal peremptorily with them in ways that eroded institutional safeguards to individual liberty. The resulting divisiveness came on top of other divisive forces in American society arising from the Vietnam War and the Civil Rights Movement.

Maintenance of the effective dispersion of power called for in the Constitution requires avoidance of the politics of scapegoating. The reason is that the politics of scapegoating creates a demand for a strong leader who will loosen institutional constraints so as to be able to deal decisively with the offending minority. People tend to rationalize impersonal, threatening forces seemingly beyond their control by blaming them on the actions of identifiable minority groups. The rise of inflation in the late 1960s appeared as such a force. Inflation added to the forces in the early 1970s that caused the middle class to feel besieged by hostile forces beyond its control. The resulting appeal of the politics of wage and price controls encouraged a majority to blame a minority for inflation.

The origin of inflation lay in the high rate of growth money created by the Fed. The political system, however, desirous of using monetary policy to promote rapid growth, found it advantageous to exploit the public's presumption that private parties create inflation through the exploitation of monopoly power. That presumption led to a demand for wage and price controls. The controls, imposed on August 15, 1971, carried the message that to control inflation

government had to prevent organized groups from selfishly demanding unreasonable shares of national income. The controls allowed government to make scapegoats of organized labor, large corporations, and sectors of the economy with politically sensitive prices such as food processors.

The combination of inflation with a fixed exchange rate created an overvalued dollar, trade deficits, and a demand for protectionism. The political system then took advantage of worker resentment over imports and fears of job losses to scapegoat foreigners. In the early 1970s, criticism of Japan became a staple of the American political scene. To provide political balance for wage controls, which were unpopular with leaders of organized labor, the Nixon Administration imposed a surcharge on imports. Although popular with labor, the unilateral adoption by America of an openly protectionist trade measure threatened to precipitate a trade war and reversed the long-standing American support for a multilateral system of world trade.[4]

The price controls of the 1970s were an extreme manifestation of the social costs of inflation. However, any time there is inflation, there is a political

[4] The complexity of economic activity made inevitable the exercise of discretion in applying the wage and price controls. That discretion allowed the arbitrary exercise of power by government bureaucracy over individuals. During the initial freeze following the announcement of controls, even individuals selling personal items in yard sales had to document the sale price of similar items to avoid transgressing the law.

Because of the difficulty of policing a large complex economy and because of the need to maintain political support, controls required widespread popular support. Popular support required the appearance of fairness. For this reason, the controls entailed all kinds of destructive intervention in the economy unrelated to the behavior of the price level. For example, unions demanded controls on interest and dividends. The general public demanded limitation on profit margins. The government used the threat of IRS audits of profit margins to influence the price setting behavior of corporations. The ideal in a constitutional democracy of limited government power enforced through due process disappeared in the populist clamor to deal with powerful corporations and unions.

Political pressures existed to make the controls permanent. The controls did not become permanent in part because of the hostility of the Ford Administration and in part because they did not work to prevent inflation. Without of course intending to, the Fed discredited them by pursuing an inflationary monetary policy. If it had not, however, the public could have seen the controls as working and they might have become a permanent feature of government control of large corporations. Price controls on oil, which lasted through 1981, almost became permanent. Those controls created an energy "shortage" and prevented America from producing the oil that would have broken the OPEC price cartel. The resulting artificially high price of oil exercised an important influence on international politics. American price controls meant that the Soviet Union, one of the largest world producers of oil, and the oil-producing countries in the Middle East gained the resources necessary to exercise significant power on the world scene.

Price controls criminalize socially useful activity. Inevitably, more and more individuals find themselves breaking the law and staying out of the criminal system only through the forbearance of a price control bureaucracy. The government can prosecute a large fraction of the population at will or in response to political pressures to deal with an unpopular group. Controls erode the general acceptance of law that allows a free society to function with a minimum of state coercion. The average citizen is law abiding because he assumes most other citizens are also law abiding. However, controls create the opposite presumption, that is, everyone else is breaking the law.

incentive to impose attenuated forms of controls through government price fixing. That incentive arises from pressures on the political system to find ways to redistribute wealth that do not require explicitly voted taxes. Although the effects of inflation interacting with price controls in individual markets are of a different magnitude from general controls, the corrosive effects are the same. Individual vice (circumventing the laws that fix prices) becomes public virtue. And respect for the rule of law erodes. (See Friedman in U.S. Congress, 6/21/73, p. 136.)

Bringing Monetary Arrangements into a Constitutional Framework

The contribution of inflation to the polarization of society and politics in the 1970s argues for bringing monetary policy into the constitutional framework in a way that limits the ability of the central bank to inflate. Congress has not established guidance for what constitutes desirable behavior of the price level. Although the *Federal Reserve Act* instructs the Fed to pursue "stable prices," it also instructs the Fed to pursue "maximum employment" (*Federal Reserve Act,* Section 2A). In practice, the absence of any instruction on how to combine the pursuit of both these objectives has meant the lack of a meaningful mandate for the behavior of the price level. The rise in the price level by a factor greater than six between 1950 and 1995 amply demonstrates the ineffectiveness of the formal mandate.

A congressional mandate to the Fed for price stability would provide an institutional safeguard against a recurrence of the divisive experience with inflation in the 1960s and the 1970s. The more clearly stated the mandate, the more understandable it would be to the public. Also, public opinion rallies more easily behind a simple mandate. The clarity of a mandate to stabilize the price level as opposed to the vagueness of a mandate to target a "low" rate of inflation means that the former would be more likely than the latter to become a permanent part of U.S. institutional arrangements.

5. SEIGNIORAGE AND THE CONSTITUTIONALLY MANDATED BUDGET PROCESS

Seigniorage as a Nonlegislated Tax

Money creation possesses implications not only for inflation, but also for the level of taxation by transferring resources to the government.[5] Moreover, money

[5] For example, in 1974 (quarterly average CPI) inflation was 12.1 percent. Hetzel (1990, p. 53) estimates that federal government revenue was 17 percent higher than it would have been with price stability. Much of the increased government revenue derived from the lack of indexing of the tax code for inflation. The personal income part of the tax code was indexed for inflation in 1985. By reducing the incentive of government to inflate, that indexing constituted an important institutional arrangement protecting against future inflation.

creation allows government to obtain resources without imposing an explicit tax. Milton Friedman (1978, p. 27) wrote:

> Since time immemorial, the major source of inflation has been the sovereign's attempt to acquire resources to wage war, to construct monuments, or for other purposes. Inflation has been irresistibly attractive to sovereigns because it is a hidden tax that at first appears painless or even pleasant, and, above all, because it is a tax that can be imposed without specific legislation. It is truly taxation without representation.

The seigniorage from money creation has implications for the U.S. constitutional system of limited government because of its potential for removing fiscal policy from the recorded budget voted on by Congress. The way the central bank handles seigniorage raises fundamental constitutional issues about openness and the separation of powers in government.

As part of the separation of powers, the Constitution assigns to Congress the power to tax and appropriate funds. As an institutional safeguard to keep fiscal policy in the hands of Congress, the Founding Fathers assigned control over the specie content of currency, and thus seigniorage, to Congress. Institutional arrangements legislated in the *Federal Reserve Act* also attempt to assure that the government will not use money creation to evade the constitutional requirement that Congress vote explicitly on taxes and appropriations.

These latter arrangements possess two aspects. One is to prevent the political system from making use of money creation as an unlegislated tax. The *Federal Reserve Act* (Section 14: Open Market Operations, (b)(1)) authorizes Fed banks "to buy and sell in the open market . . . any obligation . . . of the United States." The phrase "in the open market" implies that the monetization that occurs when the Fed purchases debt should be undertaken solely to advance monetary policy objectives rather than to alleviate the fiscal problems of the Treasury. Specifically, the Fed should not buy Treasury debt directly from the Treasury. It should not use the power of the printing press on request to turn government debt into money. This aspect of independence concerns the *size* of the Fed's asset portfolio, the quantity of money, and ultimately the price level.

The second aspect of institutional arrangements designed to assure the conduct of fiscal policy through the public acts of Congress concerns the Fed policy of transferring to the Treasury the income from the securities it holds (after meeting its operating expenses and paying the statutory dividend to member banks). In this way, the revenue from money creation appears as government revenue. Consequently, that revenue can only be spent as part of the regular appropriations process. Seigniorage revenue must be spent in ways that are voted on by Congress and that appear on budget. As explained below, this aspect of independence concerns the *composition* of the Fed's asset portfolio. (See also Goodfriend [1994].)

The Federal Reserve's Fiscal Powers

The Fed creates money by purchasing securities from the public. The seigniorage value to the government is measured by the reduction in the stock of securities held by the (nonFed) public or, equivalently, the increase in the stock of securities in the Fed's portfolio. For bookkeeping purposes, the value of this seigniorage is measured by the flow of interest payments paid on the securities in the Fed's portfolio. These bookkeeping arrangements are the heart of the institutional arrangements that support Fed independence. The Fed achieves budgetary autonomy from the political system by allocating to itself the amount of interest it needs to meet its expenses. The remainder counts as tax receipts of the government.

As part of the bookkeeping arrangements that buttress its independence, government accounts treat the Fed as a member of the public. In order to measure accurately the fiscal policy actions of the government, however, the balance sheets of both the Treasury and the Fed should be consolidated. The reason is that the Fed turns over to the Treasury the interest it receives on the government securities it holds (above its costs). As far as the government is concerned, interest paid on securities held by the Fed is essentially a wash. For the purposes of fiscal policy, the key implication is that it makes no difference whether the Treasury or the Fed sells a security. Either way, there is an increase in the debt on which the federal government must pay interest financed by some future increase in taxes or reduction in expenditure. In short, the Fed, like the Treasury, can take fiscal policy actions. Consider the following examples.

Examples of Federal Reserve Fiscal Policy

The Fed could make discount window loans to an insolvent bank. For example, in 1984 Fed loans to Continental Illinois Bank amounted to somewhat more than $7 billion, 85 percent of the bank's uninsured deposits. In conjunction with such lending, the Fed sells government securities from its portfolio to keep money and the price level unchanged. That is, it engages in a pure fiscal policy action with no consequences for monetary policy. Government debt in the hands of the (nonFed) public rises. Control over the composition of its asset portfolio gives the Fed the ability to engage in fiscal policy; in this case, it is in the form of credit allocation.[6] (See Goodfriend and King [1988] and Schwartz [1992].)

Consider next the direct monetization of Treasury assets that occurs when the Fed acquires assets from the Treasury's Exchange Stabilization Fund (ESF).

[6] The credit allocation arises because of the nonmarket allocation of funds. In this example, the Fed also transfers the liability for the insolvency from the uninsured depositors to the FDIC. Because the premiums that banks pay to the FDIC go into general federal government revenues and because the disbursements of the FDIC are government expenditures, the Fed transfers the liability to the taxpayer.

These assets take the form either of SDRs or foreign exchange.[7] When the Fed acquires assets from the ESF, it credits the Treasury's deposit account at the New York Fed. When the Treasury draws down its newly acquired deposits, the revenues of the banking system increase. The Fed then sells U.S. government securities out of its portfolio to offset that increase. The net result is to substitute either an SDR or an asset denominated in foreign exchange for a U.S. government security in the Fed's portfolio.[8] Government securities held by the (nonFed) public rise, and the Fed finances the activities of the ESF—foreign exchange intervention and lending to foreign countries.

As a final example, consider a swap between the Fed and a foreign central bank. For example, in a swap with the central bank of Mexico, the Fed accepts peso deposits in exchange for dollar deposits. The Fed invests the pesos in a peso-denominated security. When Mexico spends the dollars it receives, bank reserves in the United States increase. The Fed then sells a U.S. government security out of its portfolio to offset that increase. The net result is to substitute a peso-denominated security for a U.S. government security in the Fed's portfolio. Government securities held by the (nonFed) public rise. And the Fed lends funds to Mexico.[9]

---

[7] The International Monetary Fund periodically allocates to member countries special drawing rights (SDRs), which the ESF carries as assets.

[8] In 1988 and 1989, the U.S. Treasury and the Fed engaged in coordinated sterilized foreign exchange intervention with other central banks to counter strength in the dollar. In December 1987, the mark/dollar exchange rate was 1.6. In May 1988, the yen/dollar exchange rate was 125. By September 1989, the value of the dollar had risen so that 1.95 marks exchanged for one dollar and the 145 yen for one dollar. In 1989, the administration became concerned about the appreciation of the dollar given a large U.S. current account deficit. As a consequence, the administration and the Fed began sterilized purchases of marks and yen.

The Fed and the ESF divided their purchases of yen and marks. When the ESF ran out of dollars to sell, it obtained additional dollars from the Fed both through warehousing and through monetizing the SDRs it held. In 1989, the Fed's foreign-exchange-related transactions added about $23 billion to reserves. That was more than the additions to currency that year, and the Fed sold on net about $10 billion in government securities.

The SDRs on the books of the Fed increased from $5.0 billion at the end of 1988 to $8.5 billion at the end of 1989. In 1989, Fed warehousing of foreign currencies for the Treasury rose from zero to $7 billion. Fed monetization of the SDRs held by the ESF increases the ESF's assets permanently as the ESF uses the dollars it acquires to acquire interest-bearing assets. (See Schwartz [1997], especially Table 1.)

Figures on SDRs are from the Fed's balance sheet reported in the *Federal Reserve Bulletin*. Figures on Fed warehousing are from quarterly reports, "Treasury and Federal Reserve Foreign Exchange Operations," in the Federal Reserve Bank of New York *Quarterly Review.*

Broaddus and Goodfriend (1995) criticize such interventions for sending contradictory signals about the stance of monetary policy. In this case, the dollar was strengthening because the FOMC had raised its funds rate peg to almost 10 percent in May 1989 to contain a rise in inflation. By selling dollars to weaken the foreign exchange value of the dollar, the FOMC was sending an opposite message about the desired stance of monetary policy from what it was sending domestically by raising the funds rate. Kaminsky and Lewis (1996) make the same point.

[9] For example, in September 1989, Mexico drew on its swap line with the Fed for $784.1 million dollars. At the same time, it drew on an ESF swap line for $384.1 million. Figures on

The Fed established swap lines with foreign central banks in 1962 to defend the fixed exchange rate system without raising interest rates (Hetzel 1996). The collapse of the fixed rate system in spring 1973 eliminated the original rationale for swaps. The Fed, however, put them to another use. For instance, in 1973, the administration asked the Fed to help Italy deal with the increase in its balance of payments deficit in the aftermath of the large rise in oil prices. "The Federal Reserve . . . came to the aid of Italy, whose chronic political instability prevented rapid response to the energy crisis. The central bank expanded its swap line with the Bank of Italy from $2 billion to $3 billion to help that country finance imports in the short run" (Wells 1994, p. 125).

The use of swaps to provide short-term assistance to foreign countries prompted a debate within the Federal Open Market Committee. In response to a question from a governor about whether the Fed might provide long-term assistance to Italy, Chairman Burns (Board of Governors 1974, p. 783) responded:

> If the Federal Reserve were to abandon the principle that the swap lines were available only to meet short-term needs, there would be a natural tendency for other agencies of Government to look to the System, rather than to Congress, for the resources to deal with a broad variety of international financial and political problems. If the System were to provide those resources it would, in effect, be substituting its own authority for that of the Congress. A decisive case could then be made in support of the charge that the System was using Federal moneys without regard to the intent of the Congress.

Limiting the Federal Reserve's Ability to take Fiscal Actions

Congress has delegated to the Fed the right to exercise the public monopoly on the creation of money. Through creation of money, the Fed acquires a portfolio of government securities. Although this portfolio arises out of the conduct of monetary policy, it allows the Fed to undertake fiscal policy actions independent of Congress, as illustrated above.

Fiscal policy actions taken by the Fed are not subject to the open public debate generated by congressional actions. Therefore, they limit the government accountability that is encouraged by the free flow of information. The Fed should avoid fiscal policy actions not integrally related to its monetary responsibilities. A restriction that permitted the Fed to hold only government securities and to acquire them only in the open market would achieve this result.

---

swap line drawings are from quarterly reports, "Treasury and Federal Reserve Foreign Exchange Operations," in the Federal Reserve Bank of New York *Quarterly Review*.

The Fed does not lose interest on the assets in its portfolio as it receives interest on the peso-denominated assets. The basic point is that the Fed can engage in the loan transaction with Mexico because of its control over seigniorage. That is, the ability to create and extinguish fiat money allows it to purchase peso-denominated assets.

Fed independence would then complement the constitutional provision: "No money shall be drawn from the Treasury, but in consequence of appropriations made by law."

On the most general level, the issue is preservation of the constitutional safeguards that give content to popular sovereignty. Those safeguards facilitate the monitoring of government by the public. Because inflation imposes an unlegislated tax, discretionary control of inflation reduces the monitoring ability of the public. Also, the use of seigniorage revenues by the central bank for purposes other than financing its own operation reduces the public's ability to monitor government activities by limiting public discussion. If the use of the central bank's seigniorage revenues is directed by the Executive Branch, it erodes the separation of powers provided for in the Constitution.

6. CONCLUDING COMMENTS

Balancing Independence and Accountability

Central bank independence can help to prevent monetary policy from becoming subservient to fiscal policy. Preventing that subservience is an important part of facilitating the monitoring of government by the public. A central bank ultimately controls only money creation. A legislative mandate for price stability would limit political pressure to use money creation to achieve goals that may be socially desirable but beyond the reach of a central bank. Although money creation does not create real resources, it can impose a tax. A price rule would help keep government finance honest.

An independent central bank also needs to be accountable. The Fed chairman does testify regularly before congressional committees. However, there is a tension between the accountability provided by congressional oversight and Fed independence. The budgetary benefits of a strong economy have, in the past, encouraged some congressmen to pressure the Fed for low interest rates and, in effect, inflationary money growth. A mandate requiring the Fed to stabilize the price level would minimize this pressure, enhance independence, and provide a clear standard with which Congress could assess the Fed's performance.[10]

In a democracy, the legitimacy of central bank independence must rest on a public belief that the central bank is accountable. That belief derives from open debate encouraged by a transparency of the objectives of monetary policy. Independence combined with discretion impairs accountability and encourages political pressures that threaten the Fed's independence. Independence combined within a rule mandating price stability would balance independence and accountability.

[10] A mandate to stabilize the price level is not a complete rule unless accompanied by an explicit strategy. If Congress were to impose such a mandate, it would probably allow the Fed to select the strategy but require it to make that strategy explicit.

Rule of Law, Not Men

The Constitution originally provided for a commodity monetary standard. In doing so, it restricted the power of government significantly: the government could not manipulate the price level. In the twentieth century, the disappearance of the international gold standard that began with World War I led to a fiat money standard. The long, painful process of learning how to manage a fiat currency, where government sets the price level, has influenced significantly the history of the twentieth century (Friedman and Schwartz 1963; Goodfriend 1997).

At present, monetary arrangements are working well. However, monetary policy depends too much on the good luck of having wise policymakers. Widespread public support for a clear rule to guide the conduct of monetary policy would provide for a continuation of the current period of monetary stability. The major disasters of monetary policy in the twentieth century—the depression of the 1930s and the inflation of the 1970s—derived largely from a lack of public understanding that the central bank is responsible for the price level. A mandate clearly assigning responsibility for the price level to the Fed and requiring the Fed to stabilize the price level would prevent future major mistakes in monetary policy.

A rule that required the Fed to stabilize the price level and that eliminated its discretion over the use of seigniorage by requiring it to hold only government securities in its portfolio would complement the constitutional framework that constrains the exercise of government power. Such a rule would protect the public from the arbitrary redistributions of wealth that accompany unanticipated inflation and the interaction of sustained inflation with price controls. It would also prevent the political system from imposing an unlegislated tax in the form of inflation and assure that seigniorage is spent only as part of the congressional appropriations process.

The balance between rules and discretion in the exercise of power by government is the central issue in a constitutional democracy. That issue is also central to the design of a country's monetary institutions. Monetary institutions should be based on rules that are thought of as part of the broad constitutional framework. To protect property rights and the ability of citizens to monitor government, those rules should constrain the use of seigniorage and the recourse to an inflation tax.

APPENDIX : A BRIEF HISTORY OF MONEY AND THE CONSTITUTION

At the Constitutional Convention, members frequently cited the issuance of paper money by the states as a major abuse of power by government and a source of civil discord. For example, Governor Morris (Farrand 1911, Vol. 2, p. 299) talked about "the history of paper emissions . . . with all the distressing effects (of such measures) before their eyes." Madison (Farrand 1911, Vol. 1, p. 317) stated:

> He considered the emissions of paper money . . . as also aggressions. The States relatively to one another being each of them either Debtor or Creditor; The Creditor States must suffer unjustly from every emission by the debtor States. We have seen retaliating acts on this subject which threatened danger not to the harmony only, but the tranquillity of the Union.

In making the case for the checks and balances of a federal form of government, James Madison and others pointed to the overissue of paper money as an example of abuse of unrestrained government power. Madison (Farrand 1911, Vol. 1, pp. 134–36; Vol. 2, pp. 76–77) argued that the principal reasons for a national government were to provide

> . . . for the security of private rights, and the steady dispensation of justice. Interferences with these were evils which had more perhaps than anything else produced this convention. . . . All civilized Societies would be divided into different Sects, Factions, & interests, as they happen to consist of rich & poor, debtors and creditors. . . . In all cases where a majority are united by a common interest or passion, the rights of the minority are in danger. . . . We have seen the mere distinction of colour made in the most enlightened period of time, a ground of the most oppressive dominion ever exercised by man over man. . . . The only remedy is to enlarge the sphere, & thereby divide the community into so great a number of interests & parties that in the 1st place a majority will not be likely at the same moment to have a common interest separate from that of the whole or of the minority; and in the 2d. place, that in case they shd. have such an interest, they may not be apt to unite in the pursuit of it.
>
> Emissions of paper money, largesses to the people—a remission of debts and similar measures, will at sometimes be popular, and will be pushed for that reason. . . . it is necessary to introduce such a balance of powers and interests, as will guarantee the provisions on paper. Instead therefore of contenting ourselves with laying down the Theory in the Constitution that each department ought to be separate & distinct, it was proposed to add a defensive power to each which should maintain the Theory in practice.

Alexander Hamilton (Farrand 1911, Vol. 1, p. 288) argued:

> In every community where industry is encouraged, there will be a division of it into the few & the many. Hence separate interests will arise. There will be

> debtors & Creditors &c. Give all power to the many, they will oppress the few. Give all power to the few they will oppress the many. Both therefore ought to have power, that each may defend itself agst. the other. To the want of this check we owe our paper money.

Under the Articles of Confederation, Congress had the power to "emit bills of credit," that is, issue paper money. Governor Morris moved to omit that power from the powers assigned to Congress by the Constitution. Several delegates objected:

> Col Mason had doubts on the subject. . . . Though he had a mortal hatred to paper money, yet as he could not foresee all emergencies, he was unwilling to tie the hands of the Legislature. He observed that the late war could not have been carried on had such a prohibition existed.
>
> Mr. Randolph, notwithstanding his antipathy to paper money, could not agree to strike out the words, as he could not foresee all the occasions that might arise.

In opposition,

> Mr. Elseworth thought this a favorable moment to shut and bar the door against paper money. The mischiefs of the various experiments which had been made were now fresh in the public mind and had excited the disgust of all the respectable part of America. By withholding the power from the new Governt. more friends of influence would be gained to it than by almost any thing else—Paper money can in no case be necessary—Give the Government credit, and other resources will offer—The power may do harm, never good.
>
> Mr. Wilson. It will have a most salutary influence on the credit of the U. States to remove the possibility of paper money. This expedient can never succeed whilst its mischiefs are remembered. And as long as it can be resorted to, it will be a bar to other resources.
>
> Mr. Butler . . . was urgent for disarming the Government of such a power. . . . Mr. Read thought the words, if not struck out, would be as alarming as the mark of the Beast in Revelations.
>
> (Farrand 1911, Vol. 2, pp. 309–11.)

On the vote to omit "and emit bills of credit" from the Constitution, that is, to give Congress the power to issue paper money, nine of the delegates voted in favor and two voted against. In a later letter to Timothy Pickering, Robert Morris, who had been a delegate, wrote, "Propositions to countenance the issue of paper money, and the consequent violation of contracts, must have met with all the opposition I could make. . . . to the best of my recollection, this was the only part which passed without cavil" (Farrand 1911, Vol. 3, p. 419).

In 1862, Congress first authorized the issuance of paper money as legal tender (Greenbacks) as an expedient means of financing the Civil War. Timberlake (1993, p. 143) wrote:

> Up until the time of the Civil War, almost no one had seriously considered interpreting the money clauses in the Constitution in any light except that of prohibiting state and federal issues of currency on the basis of discretionary authority. "To coin money" meant to provide the technical facilities for minting coins. "Regulate the value thereof" meant only to specify a weight of fine gold or silver as equal to a number of the units of account, which were dollars.

Eventually, however, the Supreme Court decided in favor of the constitutionality of Greenbacks and the issuance by the federal government of paper money. At the time, these legal tender decisions were highly politicized (Timberlake 1993, pp. 133–45). President Grant appointed to the Supreme Court individuals who favored the constitutionality of Greenbacks as legal tender. Ultimately, the government issuance of paper money as legal tender was made inevitable by the change in the prevailing interpretation of the Constitution. The government's power to issue paper money became inevitable with the demise of the view that Congress possesses no power not expressly granted to it by the Constitution and the emergence of the view that Congress possesses all powers not explicitly denied to it.

In the final legal tender case, *Julliard v. Greenman* of 1884, the Supreme Court decided that Congress had the power to issue paper money and make it legal tender in peacetime as well as wartime (Timberlake 1993, p. 137). The issuance of paper money then ceased to be a Constitutional issue (capital C) in the sense of an issue decided by the Supreme Court. It remains, however, a constitutional issue (small c) because of the role of seigniorage as a tax and the insecurity of property rights engendered by inflation.

REFERENCES

Beloff, Max, ed. *The Federalist*. New York: Basil Blackwell, 1987.

Board of Governors of the Federal Reserve System. *Minutes of the Federal Open Market Committee.* 1974.

Broaddus, J. Alfred, Jr., and Marvin Goodfriend. "Foreign Exchange Operations and the Federal Reserve," Federal Reserve Bank of Richmond *1995 Annual Report*.

Commager, Henry Steele. *Documents of American History,* 7th ed. New York: Appleton, Century, Crofts, 1963.

Farrand, Max. *The Records of the Federal Convention of 1787,* Vols. 1–4. New Haven: Yale University Press, 1911.

Federal Reserve Bank of New York. "Treasury and Federal Reserve Foreign Exchange Operations," Federal Reserve Bank of New York *Quarterly Review,* vol. 14 (Winter 1989–90), pp. 69–74.

Feldstein, Martin. "The Costs and Benefits of Going from Low Inflation to Price Stability," in Christina D. Romer and David H. Romer, eds., *Reducing Inflation: Motivation and Strategy.* Chicago: University of Chicago Press, 1997.

The Financial Times. "Pöhl Throws a Gauntlet," January 23, 1990, p. 16.

Ford, Paul L., ed. *The Writings of Thomas Jefferson,* Vol. 9. New York: G. P. Putnam's, 1898.

Friedman, Milton. *Tax Limitation, Inflation and the Role of Government*. Dallas: The Fisher Institute, 1978.

_____. "Should There Be an Independent Monetary Authority?" in Leland Yeager, ed., *In Search of a Monetary Constitution*. Cambridge, Mass.: Harvard University Press, 1962.

_____. *A Program For Monetary Stability*. New York: Fordham University Press, 1960.

_____, and Anna J. Schwartz. *A Monetary History of the United States, 1867–1960.* Princeton: Princeton University Press, 1963.

Goodfriend, Marvin. "Monetary Policy Comes of Age: A 20th Century Odyssey," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 83 (Winter 1997), pp. 1–22.

_____. "Why We Need An 'Accord' for Federal Reserve Credit Policy," *Journal of Money, Credit, and Banking,* vol. 26 (August 1994), pp. 572–84.

_____, and Robert G. King. "Financial Deregulation, Monetary Policy, and Central Banking," Federal Reserve Bank of Richmond *Economic Review,* vol. 74 (May/June 1988), pp. 3–22.

Hetzel, Robert L. "Sterilized Foreign Exchange Intervention: The Fed Debate in the 1960s," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 82 (Spring 1996), pp. 21–46.

_____. "A Mandate for Price Stability," Federal Reserve Bank of Richmond *Economic Review,* vol. 76 (March/April 1990), pp. 45–55.

Kaminsky, Graciela L., and Karen K. Lewis. "Does Foreign Exchange Intervention Signal Future Monetary Policy?" *Journal of Monetary Economics,* vol. 37 (April 1996), pp. 285–312.

Lipscomb, Andrew A., ed. *Writings of Thomas Jefferson,* Vol. 10. Washington: The Thomas Jefferson Memorial Association, 1903.

Lucas, Robert E., Jr. "Rules, Discretion, and the Role of the Economic Advisor," in Robert E. Lucas, Jr., *Studies in Business-Cycle Theory*. Cambridge, Mass.: The MIT Press, 1983.

Rakove, Jack N. *Original Meanings, Politics and Ideas in the Making of the Constitution*. New York: Alfred A. Knopf, 1996.

Rutland, Robert A., ed. *The Papers of James Madison*, Vol. 9. Chicago: University of Chicago Press, 1975.

Schwartz, Anna J. "From Obscurity to Notoriety: A Biography of the Exchange Stabilization Fund," *Journal of Money, Credit, and Banking*, vol. 29 (May 1997) pp. 135–53.

_____. "The Misuse of the Fed's Discount Window," Federal Reserve Bank of St. Louis *Economic Review,* vol. 74 (September/October 1992), pp. 58–69.

Timberlake, Richard H. *Monetary Policy in the United States*. Chicago: University of Chicago Press, 1993.

U.S. Congress. "How Well Are Fluctuating Exchange Rates Working?" Hearings before the Subcommittee on International Economics of the Joint Economic Committee, 93 Cong. 1 Sess., June 20, 21, 26, and 27, 1973.

Wells, Wyatt C. *Economist in an Uncertain World: Arthur F. Burns and the Federal Reserve, 1970–78*. New York: Columbia University Press, 1994.

Tax Disincentives to Commercial Bank Lending

Anatoli Kuprianov

The Tax Reform Act of 1986 made sweeping changes to the U.S. tax code. It lowered statutory tax rates on both corporate and personal income while eliminating the investment tax credit and a host of other specialized tax deductions in an effort to ensure that all firms paid similar tax rates. The act devoted special attention to commercial banks. Studies commissioned by Congress had found that the commercial banking industry paid much lower average tax rates than most other firms, reinforcing a perception that banks enjoyed many unfair tax advantages. With passage of the Tax Reform Act, the industry lost many tax preferences it had previously enjoyed. Available evidence suggests that tax reform achieved its goal, at least insofar as the commercial banking industry is concerned: average tax rates paid by the U.S. banking industry rose from 24 percent in 1986 to 41 percent in 1995.

Some of the tax preferences banks lost under tax reform originally had been intended to offset the costs of implicit taxes such as the non-interest-bearing reserve requirements banks are obligated to hold with the Federal Reserve (the Fed) as well as the cost of other regulations (Neubig 1984). Under the current tax code, banks face the same treatment as all other financial intermediaries but are still subject to the aforementioned costs. Moreover, Henderson (1987) found that the cost of reserve requirements had not been offset by implicit subsidies associated with the banking charter, such as access to the discount window.

Banks have long argued that the costs of reserve requirements and other burdensome regulations put them at a competitive disadvantage relative to other financial intermediaries. This assertion has received some support from McCauley and Seth (1992), who found that foreign banks had gained a 45 percent share of the U.S. commercial and industrial loan market by 1991 and attributed this trend to the burden of reserve requirements imposed on U.S. banks.

Mike Dotsey, Jeff Lacker, and Ned Prescott provided helpful comments on earlier drafts of this article. Special thanks go to Leigh Ribble of the Board of Governors staff, who made available the data needed to complete the study. Any remaining errors or omissions are the author's. The views expressed are those of the author and do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

Deposit insurance premiums levied on U.S. banks also may be viewed as an implicit tax in certain circumstances. Congress recently enacted legislation requiring commercial banks to help pay for the recapitalization of the thrift industry's deposit insurance fund through a special deposit insurance surcharge. If they are fairly priced, deposit insurance premiums represent a cost of doing business and not a tax. But when a surcharge is imposed to fund other purposes, deposit insurance premiums may constitute a tax on banks. One interesting question, then, is how significant this tax is in relation to the overall effective tax rate on banks.

These observations raise fundamental issues about the effects of explicit and implicit taxes on the financial system. According to the U.S. Flow of Funds Accounts, the commercial banking industry's share of total credit extended in the United States has fallen steadily in recent years, from almost 45 percent of all credit market assets in 1952 to 22 percent in 1995. While financial innovation is most often blamed for this process of disintermediation, it is worth examining whether the tax and regulatory policies may have contributed to this trend.

This study takes a first step toward analyzing the burden of U.S. bank tax and regulatory policies by developing a comprehensive measure of the overall marginal effective tax rate on commercial bank intermediation. Economists have long recognized that average tax rates do not provide a good measure of the tax disincentives to investment. Most contemporary studies focus instead on the marginal effective tax rate, which measures the marginal tax on investment returns. Studies of bank taxation have been a notable exception to this rule—existing studies have sought to measure the impact of tax reform by estimating average effective tax rates. None of these studies has examined the marginal tax rate on bank lending. Such an exercise turns out to be worthwhile, as it produces some surprising results. In particular, it finds that the behavior of the average effective tax rate has not been a good indicator of the tax disincentives to commercial bank lending.

The discussion that follows begins in Section 1 with an examination of the conceptual issues associated with measuring effective tax rates. Section 2 develops a financial model of banks that can be used to estimate an effective tax rate on commercial bank intermediation. Empirical results are presented in Section 3. The final section reviews the conclusions of the analysis.

1. MEASURING EFFECTIVE TAX RATES ON CAPITAL INCOME

An effective tax rate is a summary statistic that measures the tax burden associated with an activity. Tax codes stipulate not only a statutory tax rate, but also a set of rules for calculating taxable income. These rules often do not yield a

measure of true economic income, however. For this reason, the effective tax rate on investment can differ substantially from the statutory tax rate.

Effective tax rates are sometimes used to measure the impact of taxes on incentives. Many studies compute an average effective tax rate, defined as actual taxes paid divided by capital income, to provide a summary statistic of the tax burden on a particular firm or industry.[1] Fullerton (1984) gives several reasons why average effective tax rates may not accurately measure the disincentives to investment created by the tax code, however. First, relying solely on U.S. taxes paid by a corporation ignores foreign taxes paid by the firm. Second, profits measured for tax purposes differ from profits measured for financial reporting. Third, taxes paid in a given year might not be related to actual profits earned that year due to carryforwards of previous losses and tax credits. Finally, profit measures typically used for calculating average effective tax rates are broken down by firm or by industry rather than by asset class. Thus, while average effective tax rates may be appropriate for measuring cash flows and distributional burden, they do not necessarily measure the disincentives to investment inherent in the tax code.

More recent research on the incentive effects of taxation has focused on the "marginal effective tax rate," which measures the extra tax resulting from a hypothetical marginal investment by a firm in a given industry. A marginal effective tax rate measures the wedge between the marginal social return to a capital asset and the rate of return earned by the investors who finance its purchase. This wedge can be viewed as a measure of the disincentives to investment created by the tax code.

Thus, marginal effective tax rates are better suited to capture disincentives to investment. Moreover, average tax rates do not reflect the burden of implicit taxes such as reserve requirements or deposit insurance surcharges.[2] When Fullerton and Henderson (1985) compare average and marginal effective tax rates for 18 industries, they find almost no correspondence between the two measures.

The effective tax rate methodology can be extended to include personal taxes paid on dividends and interest as well as corporate taxes. Measures of the effective tax rate that include personal income taxes can be used to analyze the effect of taxation on the intertemporal allocation of resources. If one is interested in the allocation of capital among competing uses, however, or among competing firms engaged in similar activities but subject to different tax treatment—such as comparison of the tax disincentives to lending between

[1] Harberger's (1966) classic article on the efficiency effects of capital income taxes uses this approach.

[2] A notable exception here is Henderson (1987), who incorporates the cost of various implicit taxes into a comprehensive measure of average effective tax rates.

commercial banks and other financial intermediaries—then consideration of personal taxes is unnecessary.[3]

As noted in the introduction, previous studies on the taxation of commercial banking have focused on the impact of changes in tax laws on total taxes paid by commercial banks.[4] While such studies can be useful in evaluating the tax burden borne by the industry, the foregoing discussion suggests that they may not be useful in measuring the disincentives to traditional forms of bank credit intermediation created by corporate taxes and reserve requirements. This study differs from other studies on the taxation of commercial banking in that it estimates the marginal effective tax rate on commercial bank intermediation. Using this model, one can also estimate the disincentives to commercial bank intermediation inherent in regulations such as reserve requirements or deposit insurance surcharges. Since the analysis focuses only on the distributional impact of explicit and implicit taxes, it ignores personal taxes paid by households.

Measurements of marginal effective tax rates are typically derived using the "user cost of capital" methodology developed by Hall and Jorgenson (1967). Hall and Jorgenson's measure of user cost reflects not only the financial cost of capital—that is, the cost of financing an investment—but also the cost of depreciation expenses and income taxes. The user cost is sometimes called an implicit rental rate because it reflects the rental cost the owner of a capital asset would have to charge to cover the costs of financing the purchase of the asset along with the cost of depreciation and income taxes. In a perfectly competitive market, this user cost would exactly equal the rental rate on capital—hence the term implicit rental rate. Once the user cost of capital is derived, the marginal effective tax rate can be calculated from the difference between the before-tax return on investment and the after-tax rate of return earned by the investors who financed the investment.[5] The following section reviews this methodology in more detail.

2. TAXES, RESERVE REQUIREMENTS, AND THE COST OF CAPITAL

Based on the foregoing discussion, the measurement of effective tax rates requires a precise measure of how explicit and implicit taxes affect the user cost

[3] If capital markets are efficient, the opportunity cost of funding will not depend on which agent in the economy buys the bonds or equity issued by the firms. In this case, the cost of capital to firms does not depend on the distribution of personal tax rates. See Fullerton (1984) for a more comprehensive discussion.

[4] See, for example, Henderson (1987), O'Brien and Gelfand (1987a, b), and Neubig and Sullivan (1987). The latter three studies estimate the impact of the Tax Reform Act of 1986 on after-tax bank profits.

[5] See Bradford and Fullerton (1981) for a detailed discussion of the conceptual issues involved in measuring marginal effective tax rates.

of capital. Understanding the impact of taxes on the cost of capital requires an understanding of the basic theory of capital budgeting. Accordingly, the analysis that follows first reviews capital budgeting theory and then applies the capital budgeting model to commercial banks. This model is used to derive an expression for the cost of capital that incorporates the effects of reserve requirements as well as deposit insurance premiums and corporate income taxes.

Review of the Basic Capital Budgeting Model

To begin, consider the simple case of a firm that finances a capital investment k by issuing interest-bearing debt, D, and equity, S. Capital invested by the firm earns a constant and known rate of return of ψ per period, so gross revenues accruing to a capital stock k are ψk. Assume capital depreciates at a constant geometric rate δ. Then, the firm can only maintain its capital stock at a constant level k by investing an additional δk units of capital in each period. By doing so, the firm maintains a constant net cash flow of

$$X = (\psi - \delta)k \tag{1}$$

in perpetuity.

The value of the firm's future cash flows is determined by the cost of capital, which, in turn, is determined by the rate of return demanded by investors in capital markets. For simplicity, assume the firm's debt takes the form of a bond issued in perpetuity that pays a fixed coupon R in each period. Let ρ_1 denote the interest rate demanded by bondholders. Then, the value of the firm's outstanding debt is just the present value of all future interest payments:

$$D = \int_0^\infty e^{-\rho_1 t} R dt$$

$$= R/\rho_1. \tag{2}$$

Now suppose that all returns net of investment and interest expense are paid to shareholders as dividends, denoted E. Then

$$X = R + E. \tag{3}$$

Since both X and R are constant over time, so is E.

Let ρ_2 denote the rate of return demanded by shareholders. Then, the value of the firm's equity shares will be determined by the present value of all future dividends discounted at the rate ρ_2:

$$S = \int_0^\infty e^{-\rho_2 t} E dt$$

$$= E/\rho_2. \tag{4}$$

The value of all outstanding claims against the firm is just

$$\begin{aligned} V &= D + S \\ &= R/\rho_1 + E/\rho_2 \\ &= \rho^{-1}(R + E), \end{aligned} \tag{5}$$

where

$$\rho = \lambda_1\rho_1 + \lambda_2\rho_2, \text{ and}$$

$$\begin{aligned} \lambda_1 &= D/V \\ \lambda_2 &= S/V. \end{aligned} \tag{6}$$

The variable ρ represents the financial cost of capital. It is the rate of return the firm must earn on its investment to be able to pay the rates of return demanded by its bondholders and shareholders.[6]

Substituting from (1) and (3) into (5) yields an expression for the value of the firm in terms of its capital stock, k, and the other variables of the model:

$$V = \rho^{-1}(\psi - \delta)k. \tag{7}$$

Assuming that capital markets are perfectly competitive, the equilibrium present value of cash flows from the investment will just equal the purchase price of the capital acquired by the firm. In equilibrium, then,

$$V = k. \tag{8}$$

From equation (7), this requirement translates into the condition

$$\psi = \rho + \delta. \tag{9}$$

Equation (9) simply shows that in equilibrium the marginal rate of return on investment, ψ, will equal the sum of the financial cost of capital, ρ, which represents the rate of return required by investors, and the marginal cost of depreciation, δ. The term of the right-hand side of equation (9) is the user cost, or implicit rental rate on capital. Note that the stationary nature of this model environment ensures that λ_1 and λ_2 are both constant over time with

$$D = \lambda_1 k, \text{ and}$$

$$S = \lambda_2 k. \tag{10}$$

[6] In a more rigorously articulated model, ρ_1 and ρ_2 would differ because of varying degrees of risk associated with each type of asset. For purposes of this analysis, however, I adopt the approach common in most intermediate finance textbooks and simply assume that rates of return on various assets can differ without explicitly modeling uncertainty.

This last result, while not important to the foregoing analysis, will be useful later on.

As an aside, the well-known Modigliani-Miller Theorem states that the financial cost of capital, ρ, is independent of the firm's capital structure when capital markets are perfect—that is, when capital markets are perfectly competitive, transactions costs are negligible, and investors have as much information about the firm's investment opportunities as its managers. Under these assumptions, a firm's investment decisions are unaffected by the mix of debt and equity it issues. This result no longer holds when corporate income taxes are introduced into the model, however.

Corporate Income Taxes and the Cost of Capital

The U.S. tax code defines taxable income as operating revenues less interest, allowable depreciation, and other operating expenses. Since this analysis focuses on the effective tax rate on capital, it will abstract from any expenses not directly affecting the cost of capital or the treatment of capital-related expenses such as depreciation allowances. As before, let ψk denote gross revenues and assume that the firm maintains a constant, fixed capital stock. Let Z denote the nominal depreciation allowance permitted under the tax code. Then, taxable profits can be expressed as

$$\pi = \psi k - R - Z, \tag{11}$$

where R denotes nominal interest payments.

Let θ denote the corporate income tax rate. Then, net after-tax cash flow can be calculated by subtracting corporate income taxes from net pre-tax cash flow, as defined in equation (1):

$$X_a = (\psi - \delta)k - \theta\pi. \tag{12}$$

Combining (11) and (12) yields

$$X_a = (1 - \theta)\psi k + \theta R - \delta k + Z. \tag{13}$$

Examine the term on the right-hand side of (13). Because interest expense affects taxable income, the variable R now appears in the expression for net cash flow. As a result, the firm's capital structure will now influence its cost of capital, and therefore its investment decisions. To see how, consider the relationship between interest expense and the firm's capital stock. From equation (2), $R = \rho_1 D$. Together with equation (10), this implies

$$R = \lambda_1 \rho_1 k. \tag{14}$$

Now consider the tax deduction for depreciation. The taxable depreciation allowance will not necessarily equal true economic depreciation. In fact, the two will differ in most cases. The taxable depreciation allowance depends on

the rules for computing the depreciable lifetime of assets and the time path of the capital stock. For purposes of the present analysis, assume that Z can be factored as

$$Z = \zeta k, \tag{15}$$

where ζ is some constant. As will be seen later on, all depreciation allowances examined in this study can be factored into such a form.

Substituting (14) and (15) into (13) yields an expression for after-tax cash flows as a function of the steady-state capital stock, k, and the other underlying variables of the model:

$$X_a = [(1-\theta)\psi + \theta\lambda_1\rho_1 - (\delta - \theta\zeta)]k. \tag{16}$$

The after-tax value of the investment, V_a, is just the present value of its after-tax net cash flow discounted using the after-tax cost of capital:

$$V_a = \rho^{-1}X_a. \tag{17}$$

In equilibrium, the present value of the firm's cash flows will equal the cost of the initial capital stock purchased by the firm. Thus,

$$V_a = k.$$

This last relation implies

$$\psi = \gamma_p(\theta) + \frac{\delta - \theta\zeta}{1-\theta}, \tag{18}$$

where

$$\gamma_p(\theta) = \lambda_1\rho_1 + \lambda_2\left(\frac{\rho_2}{1-\theta}\right) \tag{19}$$

is the pre-tax financial cost of capital. The pre-tax cost of capital differs from the after-tax cost of capital, ρ, in that the after-tax return to equity, ρ_2, is divided by $(1-\theta)$ in (19). The best way to understand this result is to note that the presence of a corporate income tax requires the firm to earn a pre-tax rate of return on equity of $\rho_2/(1-\theta)$ so it can pay out an after-tax rate of ρ_2 to its shareholders.

Now examine the second term on the right-hand side of (18). This term reflects the cost of depreciation, net of any taxable depreciation expenses. To appreciate the economic interpretation of this term, note that

$$\frac{\delta - \theta\zeta}{1-\theta} = \delta + \frac{\theta(\delta - \zeta)}{1-\theta}. \tag{20}$$

The cost of depreciation in the presence of corporate income taxes is the true depreciation rate plus the cost of the tax distortion stemming from any differences between the true economic depreciation rate and the depreciation allowance permitted for tax purposes. In the special case where the taxable depreciation allowance exactly equals true economic depreciation (that is, when $\zeta = \delta$), the

right-hand side term in (20) reduces to the true economic cost of depreciation, δ. But when $\zeta < \delta$, the effective cost of depreciation under taxation is greater than it would be otherwise. In this case, the second term on the right-hand side of (20) shows how much the capital investment must earn at the margin to pay the added tax caused by the distortion in the tax code. Conversely, an excessively liberal depreciation allowance would effectively reduce the cost of depreciation.

Taken together, the sum appearing on the right-hand side of (18) reflects the firm's pre-tax user cost of capital. It shows how much the firm's capital investment must earn at the margin so as to pay investors in bond and equity markets the returns they expect after corporate income taxes and depreciation.

The User Cost of Capital for Commercial Bank Lending

Commercial banks are generally subject to the same tax rules as all other U.S. companies. Thus, the foregoing model of investment and capital budgeting can be applied to bank lending if the variables are interpreted differently. Instead of representing physical capital, let the variable k represent the dollar value of a portfolio of loans. Then, the marginal return on investment, ψ, can be viewed as the commercial loan rate. Under this interpretation, ψk denotes gross revenues from lending.

While bank loans do not depreciate the way physical capital does, banks do incur loan losses. Loan losses affect earnings in much the same way depreciation affects the productivity of physical capital in the model presented above: when a borrower defaults on a loan, the lender no longer receives income from that loan. Accordingly, let the variable δ now represent the fraction of a bank's loan portfolio that must be written off in each period. As before, assume δ is constant over time. Under this interpretation the variable Z can be viewed as the maximum loan loss provision permitted by the tax code. As with other types of depreciation allowances, the loan loss provision permitted by the tax code has not always equaled the true cost of loan losses.

To complete the analogy, let the variable D now denote the value of outstanding deposits. Then, the results derived above can be viewed as a first approximation of the user cost of capital for a bank. Applied to banks, however, the model omits at least two important features. The first is the implicit tax imposed by reserve requirements. The second is deposit insurance premiums.

Reserve requirements obligate banks to hold non-interest-bearing reserves in the form of vault cash or reserve accounts held with the Fed. Not all bank deposits are subject to reserve requirements. Currently, the Fed imposes a 10 percent reserve requirement only on transactions deposits—demand deposits and certain interest-bearing transactions accounts such as NOW

accounts.[7] In the past, however, the Fed imposed reserve requirements on certain classes of time deposits as well.

Thus, consider a bank that issues three types of deposits as well as nondeposit debt, such as subordinated debt and bank notes. Let

$$\begin{aligned} D_1 &= \textit{transaction deposits}, \\ D_2 &= \textit{reservable time deposits}, \\ D_3 &= \textit{nonreservable deposits}, \text{ and} \\ D_4 &= \textit{nondeposit debt}. \end{aligned} \tag{21}$$

Debt of type i pays an interest rate ρ_i, $i = 1, 2, 3, 4$. The cost of equity capital—that is, the rate of return required by the bank's shareholders—is ρ_5. Under these assumptions, the bank's nominal after-tax cost of capital is

$$\rho = \sum_{i=1}^{5} \lambda_i \rho_i, \tag{22}$$

where

$$\lambda_i = D_i/V, \qquad i = 1, 2, 3, 4, \text{ and}$$

$$\lambda_5 = S/V.$$

As before, assume that the λ_i, i =1,2, . . . , 5, are fixed and constant over time, so that each type of debt outstanding is proportional to the initial capital stock. Formally,

$$\begin{aligned} D_1 &= \lambda_1 k, \\ D_2 &= \lambda_2 k, \\ D_3 &= \lambda_3 k, \text{ and} \\ D_4 &= \lambda_4 k. \end{aligned} \tag{23}$$

Reserve requirements reduce the bank's interest-earning assets by the fraction of the deposits it is forced to hold as non-interest-bearing reserves. Let α_1 and α_2 denote the required reserve ratio on deposits of type D_1 and D_2, respectively. Then, total required reserves are $(\alpha_1 D_1 + \alpha_2 D_2)$. Total funds raised

[7] Lower reserve requirements apply to the first $52 million of transactions accounts outstanding at each bank, and this tranche changes each year depending on changes in the average amount of all transactions accounts outstanding. The present analysis ignores this low-reserve tranche.

by the bank, k, are allocated to loans, denoted by the variable b, plus required reserves. Formally,

$$k = b + \alpha_1 D_1 + \alpha_2 D_2.$$

Substituting in for D_1 and D_2 from equation (23) yields

$$k = b + (\alpha_1 \lambda_1 + \alpha_2 \lambda_2) k,$$

which can be rewritten as

$$(1 - \alpha_1 \lambda_1 - \alpha_2 \lambda_2) k = b. \tag{24}$$

Equation (24) expresses the relation between total funds raised and the amount available to be invested in loans. The term $(1 - \alpha_1 \lambda_1 - \alpha_2 \lambda_2)$ is the fraction of each dollar the bank raises that is available for investment in loans. The remainder goes to satisfy reserve requirements. Thus, nominal interest revenues are ψb, the true cost of depreciation is δb, and the taxable depreciation allowance is ζb. Using (24), interest and depreciation expenses can be expressed as a function of the capital stock, k. The result is

$$\begin{aligned} \psi b &= (1 - \alpha_1 \lambda_1 - \alpha_2 \lambda_2)^{-1} \psi k, \\ \delta b &= (1 - \alpha_1 \lambda_1 - \alpha_2 \lambda_2)^{-1} \delta k, \text{ and} \\ Z &= (1 - \alpha_1 \lambda_1 - \alpha_2 \lambda_2)^{-1} \zeta k. \end{aligned} \tag{25}$$

Banks are also required to pay deposit insurance premiums on all domestic deposits. Let the symbol β denote the deposit insurance premium. Then, total deposit insurance premiums paid by the bank are

$$\beta \sum_{i=1}^{3} D_i.$$

Total taxable profits are gross revenues from lending less deposit insurance premiums, interest expense, and the provision for loan losses. Letting π denote taxable profits once again and R denote total interest payments made by the bank to depositors and bondholders,

$$\pi = \psi b - \beta \Big(\sum_{i=1}^{3} D_i \Big) - R - Z. \tag{26}$$

The bank's after-tax cash flow is just its revenues less loan losses, deposit insurance premiums, and taxes:

$$X_a = (\psi - \delta) b - \beta \sum_{i=1}^{3} D_i - \theta \pi. \tag{27}$$

Substituting the expression for taxable profits (equation [26]) into (27) yields

$$X_a = (1 - \theta) \psi b - (1 - \theta) \beta \sum_{i=1}^{3} D_i + \theta R - (\delta b - \theta Z). \tag{28}$$

Consider the relation between interest expenses and the capital stock. Let R_i, $i = 1, \ldots, 4$, denote total interest payments on debt of type i. Then,

$$R_i = \rho_i D_i$$

$$= \lambda_i \rho_i k, i = 1, \ldots, 4,$$

and

$$R = \sum_{i=1}^{4} \rho_i D_i$$

$$= \rho_D k, \tag{29}$$

where

$$\rho_D = \sum_{i=1}^{4} \lambda_i \rho_i \tag{30}$$

is the weighted-average nominal interest cost.

Substituting from (25) and (29) into (28) yields an expression for net cash flows as a function of the value of the initial investment, k:

$$X_a = \Big[(1-\theta)\Big((1-\alpha_1\lambda_1-\alpha_2\lambda_2)\psi - \beta(\sum_{i=1}^{3}\lambda_i)\Big)$$

$$+\theta\rho_D - (1-\alpha_1\lambda_1-\alpha_2\lambda_2)(\delta-\theta\zeta)\Big]k. \tag{31}$$

The after-tax discounted value of this investment is $V_a = \rho^{-1}X_a$. As before, the bank's user cost of capital can be derived by imposing the equilibrium condition $V_a = k$. The result is

$$\psi = \frac{\gamma_p(\theta,\beta)}{1-\alpha_1\lambda_1-\alpha_2\lambda_2} + \frac{\delta-\theta\zeta}{1-\theta}, \tag{32}$$

where

$$\gamma_p(\theta,\beta) = \sum_{i=1}^{3}\lambda_i(\rho_i+\beta) + \lambda_4\rho_4 + \lambda_5\left(\frac{\rho_5}{1-\theta}\right). \tag{33}$$

denotes the pre-tax financial cost of capital.

These last two expressions are very similar to those derived in the previous case (equations 18 and 19) except that the pre-tax financial cost of capital in (33) now includes the cost of deposit insurance premiums. Thus, $\rho_i + \beta$ is the effective cost of issuing deposits of type i, $i = 1, 2, 3$, not including the cost of reserve requirements.

Notice also that the expression for the user cost of capital in (32) differs from the earlier user cost of capital presented in (18) in that the pre-tax financial cost of capital, $\gamma_p(\theta,\beta)$, is now divided by the term $(1-\alpha_1\lambda_1-\alpha_2\lambda_2)$ to reflect the cost of reserve requirements. The firm must now earn a marginal rate of return

$$\frac{\lambda_p(\theta,\beta)}{1-\alpha_1\lambda_1-\alpha_2\lambda_2} > \lambda_p(\theta,\beta)$$

after accounting for the after-tax loan loss expense, $\left((\delta - \theta z)/(1 - \theta)\right)$, to pay its depositors and shareholders the return on investment they expect.

Loan Loss Allowances

Banks report loan loss reserves, also known as provisions for loan losses, on their balance sheets as an estimate of probable future loan losses. Under generally accepted accounting principles (GAAP), any additions to loan loss reserves, termed loan loss allowances, are deducted from reported income in the period the provisions are made and not in the period when the loss actually occurs. When a bank subsequently determines that a loan is uncollectible, it reduces its loan loss reserves by the amount of the loss. Because the impact of the loan loss on earnings is taken into account when the loan loss reserve is created, the act of writing off the loan has no direct impact on income reported in that period.

For a variety of reasons, banks typically maintain loan loss reserves in excess of their expected losses for the coming year.[8] Before 1987, the tax code permitted all commercial banks to deduct loan loss allowances, up to a stipulated maximum, from taxable income. The Tax Reform Act of 1986 changed the rules for computing deductions for loan losses, however, reducing the loan loss deductions available to many banks. The discussion that follows describes the tax treatment of loan loss allowances, both before and after the Tax Reform Act.

The Tax Treatment of Loan Loss Allowances for Large Banks

Since 1987, "large" commercial banks (banks with assets over $500 million) have been permitted to deduct loan losses from taxable income only as they are recognized. Many analysts feel that this rule, known as the "specific charge-off" method, produces the most accurate measure of true economic income, as it requires banks to recognize both interest income and loan losses in the year they accrue.[9] If one accepts this argument, the current tax treatment of loan loss allowances accorded to large banks specifies a deductible loan loss allowance that equals the true "depreciation" of the loan portfolio. To model the post-1987 loan loss provision for large banks, then, set

$$Z = \delta b. \tag{34}$$

[8] See Walter (1991) for a more detailed discussion of the factors determining loan loss reserves.

[9] See, for example, Buynak (1987), Neubig (1984), and Neubig and Sullivan (1987). For a dissenting view, see Henderson (1987).

In this case, the user cost of capital given in equation (32) reduces to

$$\frac{\gamma_p(\theta, \beta)}{1 - \alpha_1\lambda_1 - \alpha_2\lambda_2} + \delta, \tag{32′}$$

where $\gamma_p(\theta, \beta)$ is as given in equation (33).

Tax Treatment of Loan Loss Allowances for Small Banks

A "small" bank (one with assets less than $500 million) can choose between the specific charge-off method and the "experience reserve" method. Under the experience reserve method, a bank may deduct additions to its bad debt reserves up to a maximum amount determined by the product of its eligible loans outstanding and a six-year moving average of its historical loan loss ratio. To see how this method works, let $\delta(t)$ denote the actual loan loss ratio experienced in year t, and $\bar{\delta}(t) = (1/6)\sum_{i=0}^{5}\delta(t-i)$ the moving-average of the current and past five years of loan loss ratios. Then, the maximum loan loss reserve (LLR) permitted in year t is

$$\text{LLR}(t) = \bar{\delta}(t)b(t),$$

where $b(t)$ denotes eligible loans outstanding in period t. The corresponding maximum loan loss allowance deduction is

$$Z(t) = \delta(t)b(t) + \big(\text{LLR}(t) - \text{LLR}(t-1)\big). \tag{35}$$

If the size of a bank's loan portfolio does not change over time, then the experience reserve method is roughly equivalent to the specific charge-off method. In the case of a bank with a growing loan portfolio, however, use of the experience reserve method has the effect of accelerating the recognition of future loan loss deductions and causes taxable income to understate true economic income (Neubig and Sullivan 1987). To simplify notation and better understand the properties of these provisions, assume that the loan loss ratio is constant over time; that is, assume that $\delta(t) = \delta$ for all t. Then, $\bar{\delta}(t) = \delta$ and

$$Z(t) = \delta\big(b(t) + \Delta b(t)\big), \tag{36}$$

where $\Delta b(t) = b(t) - b(t-1)$ is the change in eligible loans outstanding from year $(t-1)$ to year t. Clearly, this reduces to the specific charge-off method currently permitted to all banks when $\Delta b(t) = 0$.

To examine the more general case where the bank's loan portfolio may grow over time, let

$$\mu = \Delta b(t)/b(t).$$

Substituting this last result into equation (36) yields the following expression

$$Z(t) = \delta(1 + \mu)b(t).^{10} \tag{37}$$

Tax Treatment of Loan Loss Allowances before Tax Reform

Before 1987, commercial banks were permitted to choose among several different methods for calculating the taxable loan loss allowance: the experience method, the specific charge-off method (both as described above), and the "percentage method." The percentage method was similar to the experience method, except that the deductible loan loss allowance was 0.6 percent of total eligible loans outstanding. As before, let LLR(t) denote the allowable loan loss reserve in year t. Then, the allowable loan loss allowance would be calculated as in equation (35) above, except that in this case

$$\text{LLR}(t) = 0.006b(t).$$

Substituting this last specification into equation (35) yields the result

$$Z(t) = \delta b(t) + 0.006\Delta b(t). \tag{38}$$

In the special case where the size of a bank's loan portfolio stays constant over time, $\Delta b(t) = 0$ and the above expression reduces to $Z(t) = \delta b(t)$, which is the same as the deduction permitted under the specific charge-off method. In the more general case where $\Delta b(t) = \mu b(t)$, one obtains

$$Z(t) = (\delta + 0.006\mu)b(t). \tag{39}$$

Substituting this last result into equation (32) yields the following expression for a bank's user cost of capital under the percentage method

$$\frac{\gamma_p(\theta, \beta)}{1 - \alpha_1\lambda_1 - \alpha_2\lambda_2} + \left(\delta - \frac{0.006\theta\mu}{1 - \theta}\right), \tag{32$''$}$$

where, as before, $\gamma_p(\theta, \beta)$ is as given in equation (33).

Loan Loss Reserve Recapture Provisions

In addition to eliminating the percentage method, the Tax Reform Act of 1986 also required large banks to recapture any existing loan loss reserves in excess of actual losses. Under this provision, large banks were required to report as income a fraction of 10 percent of excess bad debt reserves in 1987, 20

[10] The astute reader will note what seems to be a logical inconsistency here, as the foregoing analysis has assumed a constant loan portfolio size while the depreciation rules allow for a growing loan portfolio. Interested readers are invited to verify that the results presented in the text would remain unchanged in all substantive respects if loan portfolio growth were taken into explicit account in the capital budgeting problem.

percent in 1988, 30 percent in 1989, and 40 percent in 1990.[11] Assuming that banks knew that their excess loan loss reserves would be subject to recapture after 1986, these recapture provisions would effectively reduce the value of the 1986 loan loss deduction by the expected present value of future excess tax payments.[12]

To calculate the present value of the recapture provisions, one must take into account any expected future changes in the statutory tax rate. In addition to mandating the recapture of the loan loss reserve, the Tax Reform Act also lowered the statutory corporate tax rate from 46 percent in 1986 to 40 percent in 1987 and to 34 percent thereafter. As a result, a dollar in loan loss reserves deducted before 1987 produced a 46-cent reduction in taxes, while the subsequent recapture of a dollar in loan loss reserves increased future taxes by a smaller amount. Thus, the present value of 1987 taxes attributable to the loan loss recapture would have been $(0.40)(0.1)e^{-\rho}$. Similarly, the present value of taxes due to the loan loss recapture for subsequent years would have been

$$(0.34)\left(0.2e^{-2\rho} + 0.3e^{-3\rho} + 0.4e^{-3\rho}\right).$$

On net, then, taking account of the recapture provisions, the expected present value of the loan loss allowance to a large bank in 1986 would have been

$$Z(1986) = \left\{\delta + 0.006[\mu - (40/46)\left(0.1e^{-\rho}\right) - (34/46)\left(0.2e^{-2\rho} + 0.3e^{-3\rho} + 0.4e^{-4\rho}\right)]\right\} b(1986). \qquad (40)$$

3. THE MARGINAL EFFECTIVE TAX RATE ON COMMERCIAL BANK LENDING

The foregoing analysis has been almost entirely theoretical, focusing on the qualitative effects of explicit and implicit taxes on the user cost of capital. A purely theoretical analysis does not permit one to gauge the quantitative importance of specific tax rules, however. Nor can it answer questions regarding

[11] The act provided exceptions for financially troubled institutions, which were permitted to defer payment of taxes on the amount of the recapture. It also permitted banks to accelerate the recapture. This last provision permitted banks reporting losses between 1987 and 1990 to avoid paying at least part of the tax on the recapture (see U.S. Congress, 1987, pp. 549–57). The present analysis ignores such considerations.

[12] Even in the absence of the Tax Reform Act of 1986, banks' authorization to use the percentage method would have expired after 1987 (Henderson 1987). The tax reform of 1969 had instituted a gradual reduction of the maximum limit on loan loss reserve deductions, and the expiration of the authority to use this method was expected to trigger some type of recapture. Nor did the banking industry have reason to expect that forthcoming legislation would reinstate this deduction. The U.S. Treasury had given the treatment of bad debt reserves special attention during the debate over tax reform (see Neubig [1984]). Therefore, although the Tax Reform Act was not passed until the summer of 1986, it seems reasonable to assume that commercial banks expected they would be required to recapture their excess loan loss reserves after 1987.

the overall impact of legislation such as the Tax Reform Act of 1986, which lowered statutory tax rates while imposing offsetting reductions in the allowable deduction for loan losses. To answer such questions, one needs an empirical measure of the user cost of capital for banks.

Figure 1 depicts the behavior of the pre-tax and after-tax user cost of capital for commercial bank lending from 1986 to 1995. This period is an interesting one, as it includes a major change in tax laws and two separate instances where reserve requirements were reduced. The data used to compute these series were obtained from various reports that all insured banks must file routinely with the federal regulatory agencies. Both series were obtained by aggregating year-end data on all domestic commercial banks.[13] As such, these series represent industrywide weighted averages. The values of the parameters characterizing tax rules and reserve requirements during this period are summarized in Table 1.

The pre-tax cost of capital in Figure 1 is calculated using the formulas specified in equations (32) and (33). In accordance with the earlier discussion of the tax treatment of loan loss allowances, the loan loss deduction for 1986 includes the present value of the future loan loss reserve recapture mandated by the Tax Reform Act, as characterized in equation (40).[14] The loan loss allowance for subsequent years is based on the formula in equation (34). Thus,

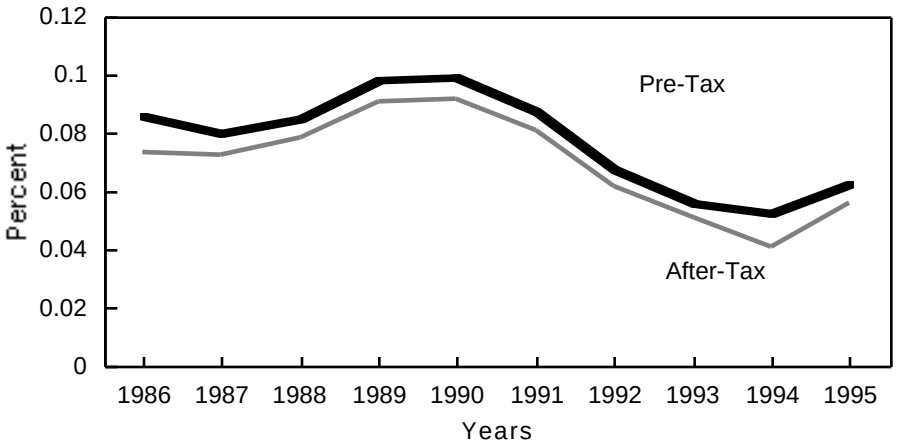
$$\zeta = \begin{cases} \delta + 0.006\big[\mu - (40/46)0.1e^{-\rho} - (34/40)\big(0.2e^{-2\rho} \\ \quad + 0.3e^{-3\rho} + 0.4e^{-4\rho}\big)\big] \text{ for } t = 1986, \text{ and} \\ \delta \text{ for } t \geq 1987. \end{cases}$$

The after-tax cost of capital is just the pre-tax cost of capital with all tax parameters, α_i, θ, and ζ, set to zero. How best to treat deposit insurance premiums presents certain conceptual problems. To the extent that deposit insurance reduces funding costs for banks, the deposit insurance premium, β, just reflects the offsetting cost of the financial guarantee. If deposit insurance were privately provided and supplied at a market-determined price, the deposit insurance premium would not be viewed as a tax on commercial bank intermediation. As noted earlier, however, FDIC deposit insurance premiums may not always reflect the fair market value of the underlying guarantee. If they are set too low, they represent a subsidy. If they are increased to raise funds for other purposes, such as rescuing a competing deposit insurance fund, deposit insurance premiums can constitute a tax. For the present, assume that deposit

[13] A more detailed description of data sources and calculations can be found in the Appendix.

[14] For now, I assume that all banks are "large" banks that are subject to the specific charge-off method and the recapture of loan loss reserves. In later sections, I will examine the marginal impact of the loan loss recapture provisions of the Tax Reform Act of 1986 and the marginal tax benefit of the experience reserve method, which small banks continued to enjoy throughout the period under consideration.

Figure 1 Pre-Tax and After-Tax User Cost of Capital



insurance is fairly priced. Accordingly, the after-tax cost of capital in Figure 1 is calculated according to the formula

$$\gamma_a(\beta) = \left[\lambda_1(\rho_1 + \beta) + \lambda_2(\rho_2 + \beta) + \lambda_3(\rho_3 + \beta) = \lambda_4\rho_4 + \lambda_5\rho_5\right] + \delta. \quad (41)$$

The marginal impact of deposit insurance premiums on the cost of capital will be examined in a later section.

The Taxation of Commercial Banking: 1986–1995

There are several ways in which one can measure the marginal effective tax rate. The simplest measure is the difference between the pre-tax and after-tax cost of capital, which reflects the marginal cost of taxes on investment returns. Alternatively, the marginal effective tax rate can be expressed as a percentage either of the pre-tax or after-tax cost of capital.[15] Figure 2 depicts the behavior of the marginal effective tax rate, measured as the difference between the pre-tax and after-tax user cost of capital. Notice that the marginal effective tax rate has fallen on average over the period in question, from a high of 126 basis points in 1986 to under 70 basis points in recent years. Most of this decline took place in the two years immediately following enactment of the Tax Reform Act of 1986, corresponding to the period in which the reductions in the statutory tax rate mandated by the act were phased in. By 1988, the marginal effective tax

[15] See Bradford and Fullerton (1981) for an analysis of the properties of these different summary statistics.

Table 1 Summary of Tax Parameter Values

	Statutory Tax Rate (θ) (percent)[a]	Present Value of the Deduction for Loan Loss Allowances ($Z(t)$), (Large Banks)	Reserve Requirements[b]
1986	46	$\delta + [0.006(\mu - \left(\frac{40}{46}\right)(0.1e^{-\rho})$	Transaction: $\alpha_1 = 12\%$
		$-\left(\frac{34}{46}\right)(0.2e^{-2\rho} + 0.3e^{-3\rho} + 0.4e^{-4\rho})b(t)]$	Time: $\alpha_2 = 3\%$[c]
1987	40[d]	$\delta b(t)$	Transaction: $\alpha_1 = 12\%$ Time: $\alpha_2 = 3\%$[c]
1988	34	$\delta b(t)$	Transaction: $\alpha_1 = 12\%$ Time: $\alpha_2 = 3\%$[c]
1989	34	$\delta b(t)$	Transaction: $\alpha_1 = 12\%$ Time: $\alpha_2 = 3\%$[c]
1990	34	$\delta b(t)$	Transaction: $\alpha_1 = 12\%$ Time: $\alpha_2 = 3\%$
1991	34	$\delta b(t)$	Transaction: $\alpha_1 = 12\%$ Time: $\alpha_2 = 0\%$
1992–1995	34	$\delta b(t)$	Transaction: $\alpha_1 = 10\%$[e] Time: $\alpha_2 = 0\%$

[a] Before the Tax Reform Act of 1986, corporations with taxable income below $100,000 were subject to a lower rate. In addition to lowering the statutory tax rate to 34 percent, the act changed the graduated tax structure. Starting in 1987, the threshold for lower tax rates was lowered to $75,000. Both before and after the Tax Reform Act, corporations with incomes exceeding the threshold were subject to a surcharge meant to recover the benefit of lower tax rates on income below the threshold. Currently, corporations must pay a 5 percent surcharge on income over $100,000 up to a maximum of $11,750. As a result, corporations with taxable incomes over $335,000 pay both an average and a marginal statutory rate of 34 percent. (For more details, see U.S. Congress [1987], pp. 271–72.) In constructing the weighted-average cost of capital, it was assumed that all banks were subject to the maximum statutory tax rate.

[b] The Garn-St. Germain Depository Institutions Act of 1982 required that $2 million of reservable liabilities of each depository institution be subject to a zero percent reserve requirement. The act instructs the Board of Governors to adjust the amount of reservable liabilities subject to this zero percent reserve requirement each year by 80 percent of the annual percentage increase in the total reservable liabilities of all depository institutions. In 1996, this zero-reserve tranche was raised to $4.3 million.

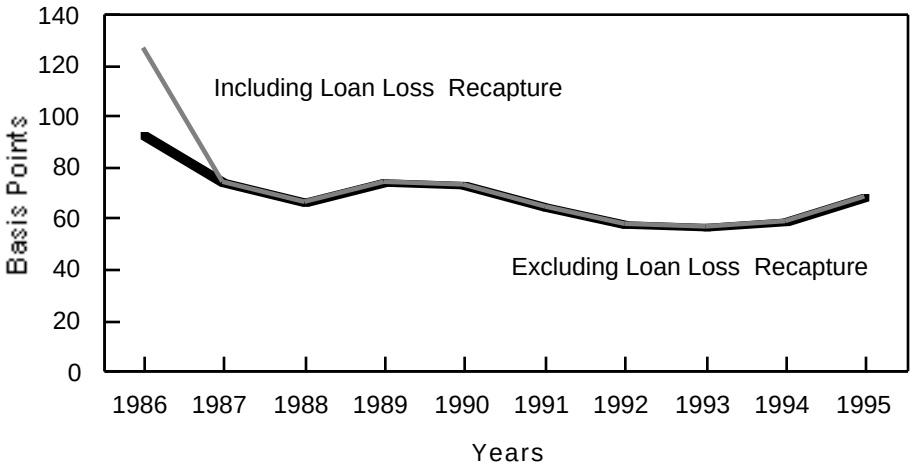
The Monetary Control Act of 1980 established a low-reserve tranche against which a 3 percent reserve requirement is applied. In 1995, a 3 percent reserve requirement was applied to the first $52 million in reservable deposits. As with the zero-reserve tranche, this amount is adjusted each year by 80 percent of the total percentage increase in the total reservable liabilities of all depository institutions. The user cost of capital calculations ignores the zero- and low-reserve tranches, since, at the margin, virtually all banks are subject to the higher reserve requirement listed in the table.

[c] During this period, reserve requirements on time deposits applied only to nonpersonal time deposits with an original maturity less than $1\frac{1}{2}$ years. The reserve requirement on nonpersonal time deposits was reduced to zero at the end of 1990.

[d] The Tax Reform Act of 1986 reduced the maximum statutory tax from 46 percent to 34 percent, effective for taxable years beginning on or after July 1, 1987. Income in taxable years including July 1, 1987, was subject to a blended rate. According to the methodology specified in the act, the effective statutory tax rate for the 1987 calendar year would have been calculated as (181/365)x(40%) + (184/365)x(34%) = 40%, as there were 181 days between January 1, 1987, and June 30, 1987, and 184 days between July 1, 1987, and December 31, 1987. For more details, see U.S. Congress (1987), pp. 272–73.

[e] The reserve requirement on transaction deposits was reduced from 12 to 10 percent in April 1992.

Figure 2 The Marginal Effective Tax Rate on Commercial Bank Lending

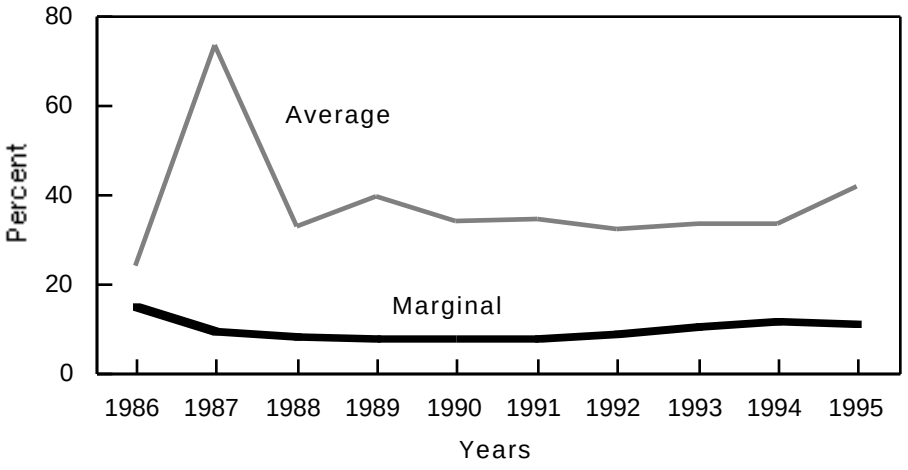


rate had fallen by almost half, to 66 basis points, suggesting that the reduction in the statutory corporate tax rate more than offset the loss of the loan loss reserve deduction.

Figure 3 compares the behavior of the marginal effective tax rate with that of the average tax rate, computed as taxes paid as a percent of pre-tax earnings. To facilitate comparison, the marginal effective tax rate in Figure 3 is expressed as a percent of the pre-tax cost of capital. In light of the earlier discussion, it should not be surprising to find that the behavior of the marginal effective and average tax rate measures differ so much. Whereas the marginal effective rate falls dramatically after 1986, the average tax rate rises. The difference in the behavior of these two series is in part due to the timing of the recognition of cash flows. Recall that the marginal effective tax rate incorporates the full present value of the future loan loss recapture in 1986—consequently, the cost of recapture does not influence the marginal tax rate in later years. In contrast, the measured average tax rate recognizes these taxes only as they accrue.

The behavior of the average tax rate also reflects other changes in tax rules that did not affect the incentive of commercial banks to make loans. One of the major provisions of the Tax Reform Act of 1986 was the repeal of the tax deductibility of interest payments on municipal bonds. Before 1987, commercial banks paid no taxes on interest earned on municipal bonds, while interest payments on bank debt issued to fund such investments was tax deductible. Partly as a result of this favorable tax treatment, the average tax rate for commercial banks tended to be low, especially when compared to the average tax rate on most other industries. The perception that banks enjoyed

Figure 3 Comparison of Average and Marginal Effective Tax Rates



too many tax advantages led Congress to repeal the tax deduction on interest from municipal bonds, along with the tax deduction for bad debt reserves.[16] As Henderson (1987), Neubig and Sullivan (1987), and O'Brien and Gelfand (1987a, b) note, however, the increase in the average tax rate due to the repeal of the tax deduction for interest on municipal bonds is largely illusory. Because interest paid on municipal bonds is not subject to federal taxes, interest rates on such bonds tend to be lower than interest rates on taxable bonds—in fact, the interest rates paid on municipal bonds tend to be comparable to after-tax interest rates on taxable bonds. Thus, the interest rate differential between taxable and nontaxable bonds can be viewed as an implicit tax. After losing this tax deduction in 1987, banks tended to substitute taxable bonds for municipal bonds, with the result that taxable income increased along with taxes paid. Although banks paid higher federal taxes on average, the impact on their after-tax return was minimal. The principal result of this change in tax laws, then, was to substitute explicit taxes paid to the federal government for implicit taxes that were previously paid to municipalities.

Despite these considerations, Figure 3 does hold a seeming puzzle. Note that the marginal effective tax rate tends to be much lower than the measured average tax rate; this despite the inclusion of the cost of reserve requirements and deposit insurance premiums in the marginal rate but not in the average rate. Three different factors can help explain this apparent anomaly. First, Henderson

[16] See U.S. Congress (1987).

(1987) notes that many large banks have substantial foreign operations, which are subject to higher tax rates. Second, to the extent that nonfinancial assets are effectively taxed at a higher rate than financial assets such as commercial loans, the result would be to raise the average tax rate above the marginal effective tax rate on lending. Finally, the marginal effective tax rate calculations derived earlier and illustrated in Figures 2 and 3 assume banks earn no pure economic profits. To the extent that banks do earn economic profits, the marginal effective tax rate on such profits just equals the statutory tax rate, which is currently 34 percent.[17]

The Long-Run Impact of Tax Reform

As noted earlier, the data depicted in Figure 2 suggest that the reductions in the statutory corporate income tax rate that took place in 1987 and 1988 more than offset the loss of the tax deduction for the bad-debt reserve. Care must be taken in interpreting this result, however, because the marginal effective tax rate is influenced by many factors, not just changes in tax rules. Moreover, the influence of the future recapture of the deduction for loan loss reserves mandated by the Tax Reform Act exerted a significant transitory influence on the marginal effective tax rate in 1986.

A measure of the long-run marginal impact of the Tax Reform Act on the marginal effective tax rate can be calculated by computing the user cost of capital for a single year under the two sets of tax rules, ignoring recapture provisions. The results of such an exercise, performed using 1986 data, are presented in Table 2. When the effect of the recapture provisions is excluded, the marginal effective tax rate for 1986 falls to 92 basis points—thus, the marginal impact of the recapture provisions was 34 basis points. Recomputing the 1986 user cost of capital assuming a 34 percent tax rate and adopting the specific charge-off method reduces the marginal effective tax rate another 30 basis points. From these two exercises, one can conclude that about half of the observed decline in the marginal effective tax rate between 1986 and 1988 was attributable to the long-run impact of tax reform. Although other factors also contributed to the observed decline in the marginal effective tax rate during this period, their influence was minimal.

Tax Reform and Small Banks

The last exercise ignored the differential treatment accorded to small banks by the Tax Reform Act and assumed that all banks lost the loan loss reserve deduction. Recall, however, that small banks were permitted to continue using the experience reserve method in determining their loan loss deduction. What impact did tax reform have on these institutions?

[17] See Bradford and Fullerton (1981) for a more comprehensive discussion of this last issue.

Table 2 The Long-Run Marginal Impact of Tax Reform on the Marginal Effective Tax Rate[a]

	Effective Tax Rate, Including Recapture (basis points)	**Effective Tax Rate, Excluding Recapture** (basis points)
Large Banks[b]		
Before Tax Reform	126	92
After Tax Reform	63	63
Small Banks[c]		
Before Tax Reform	NA	92
After Tax Reform	NA	63

[a] Calculated using 1986 year-end industrywide financial data.
[b] The marginal effective tax rate for large banks before tax reform was calculated using a statutory tax rate of 46 percent and assumes that banks used the percentage method to calculate the loan loss allowance. The tax rate after tax reform was computed assuming a 34 percent statutory tax rate and assumes that banks use the specific charge-off method.
[c] The user cost of capital for small banks is based on the same data used in the large bank example, except that the calculations assume use of the experience reserve method, under which $\zeta = \delta(t) + \bar{\delta}(t)\mu$, where $\bar{\delta}(t) = \left(\frac{1}{6}\right)\sum_{i=0}^{5}\delta(t-i)$ and μ represents the growth rate of eligible loans. As with the first exercise, the before-tax-reform user cost of capital is calculated assuming a 46 percent statutory tax rate, and the after-tax-reform user cost is computed assuming a 34 percent tax rate.

An approximate measure of the marginal effective tax rate for small banks can be obtained using industrywide weighted averages. Specifically, consider a hypothetical representative "small" bank that experienced the same realized loan loss ratios and growth rates in outstanding loans as did the industry in the aggregate. Next, compute the user cost of capital assuming that this bank takes advantage of its option to use the experience reserve method, as characterized in equation (37). This last result can then be used to compute an marginal effective tax rate measure for small banks using the experience reserve method.

Table 3 compares the marginal effective tax rates under the specific charge-off (small bank) method with that obtained using the experience reserve (large bank) method.[18] Notice that the two tax rates differ by no more than 6 basis points after 1986. The difference between the two in 1986 can be accounted for almost entirely by the present value of future loan loss recoveries imposed on large banks.

Evidently, the favorable treatment of loan loss reserves accorded to small banks by the Tax Reform Act of 1986 has had only a small impact on the

[18] For 1986, the "large" bank effective tax rate calculation assumes that banks use the percentage method to calculate the taxable deduction for loan loss reserves.

Table 3 Comparison of the Marginal Effective Tax Rate for Large and Small Banks

	Large Banks	Small Banks	Difference
1986	126	92	34
1987	74	71	2
1988	66	64	2
1989	74	71	3
1990	73	72	1
1991	64	66	−2
1992	57	58	−1
1993	57	53	4
1994	59	53	6
1995	69	63	5

Note: All figures are expressed as basis points.

marginal effective tax rate on lending. The largest benefit to small banks conferred by the act was in exempting them from the recapture of past excess contributions to loan loss reserves.

The Cost of Reserve Requirements

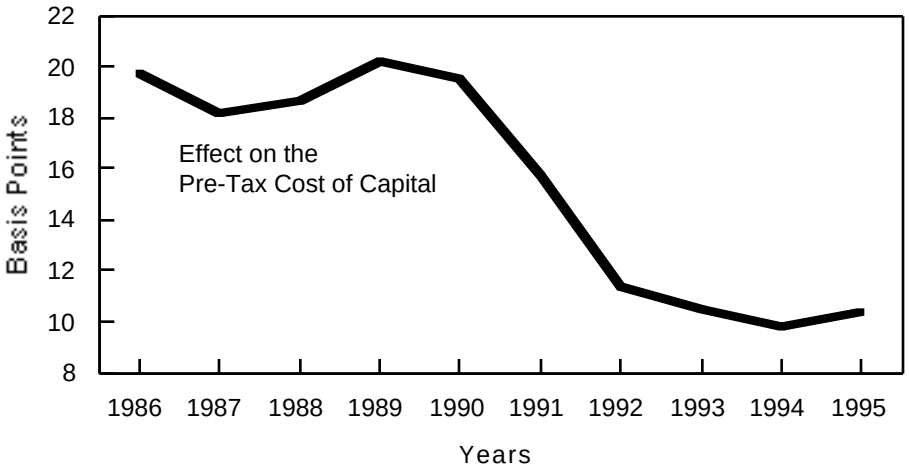
The derivation of the user cost of capital given in equation (32) showed how non-interest-bearing reserve requirements increase the cost of funding a loan. Figure 4 shows the net overall impact of reserve requirements on the user cost of capital from 1986 to 1995, obtained by calculating the difference between the pre-tax cost of capital, including the cost of reserve requirements, and the pre-tax cost of capital net of reserve requirements:

$$\frac{\gamma_p(\theta, \beta)}{1 - \alpha_1\lambda_1 - \alpha_2\lambda_2} - \gamma_p(\theta, \beta).$$

As Figure 4 shows, the cost of reserve requirements fell significantly after 1990, from approximately 20 basis points in that year to just over 10 basis points in 1994. There are at least three factors that might account for this decline. The first was the elimination of reserve requirements against time deposits in 1991 and a subsequent reduction from 12 to 10 percent in the required reserve ratio for transaction deposits in 1992. The second is a decline in interest rates. The third is a decline in banks' reliance on reservable deposits—to the extent that banks substitute nonreservable liabilities for those bearing reserve requirements, they can effectively avoid paying the implicit reserve requirement tax.

The elimination of reserve requirements against time deposits accounted for almost 2 basis points of the observed decline in Figure 4, while the reduction in the reserve ratio for transaction deposits accounted for just under 3 basis

Figure 4 The Cost of Reserve Requirements



points.[19] The remainder of the reduction can be attributed to falling interest rates. Changes in the ratio of reservable deposits to other liabilities does not appear to have contributed to the observed reduction in the overall cost of reserve requirements.[20]

Deposit Insurance and the Cost of Capital

From 1935 to 1989 all insured U.S. commercial banks paid the FDIC an annual statutory deposit insurance premium of 0.0833 percent of domestic deposits (or 8.33 basis points). The effective deposit insurance premium was often much lower, however, because the FDIC frequently rebated some portion of these premiums. Such rebates ended in the late 1980s after a large increase in the number of bank failures threatened to deplete the FDIC's Bank Insurance Fund (BIF). Using its newly acquired authority to increase deposit insurance assessments, the FDIC raised its assessments to 0.2125 percent in 1991 and again in 1992

[19] Estimated cost savings stemming from the 1991 elimination of reserve requirements against nonpersonal time deposits were obtained by measuring the marginal cost of such reserve requirements at the end of 1990. Similarly, estimated cost savings associated with the 1992 reduction in required reserve ratios on transaction deposits reflect the marginal cost of holding an extra 2 percent reserve requirement at the end of 1991.

[20] Although the importance of demand and other transaction deposits has fallen substantially in the past 25 years, transaction deposits accounted for approximately 20 percent of the value of debt plus equity throughout the period 1986–1990, falling slightly from 1985 to 1990 and rising modestly thereafter.

to 0.23 percent. Following a congressional mandate, the agency adopted risk-based assessments in 1993. Under this latter system, banks paid assessments in the range of 0.23 to 0.31 percent. The minimum assessment was lowered to 0.044 percent in mid-1995 after BIF reached its mandated capitalization level, and well-capitalized and well-managed banks received a small rebate that year. In 1996, the FDIC reduced its risk-based premiums further to a range of zero to 0.31 percent.[21]

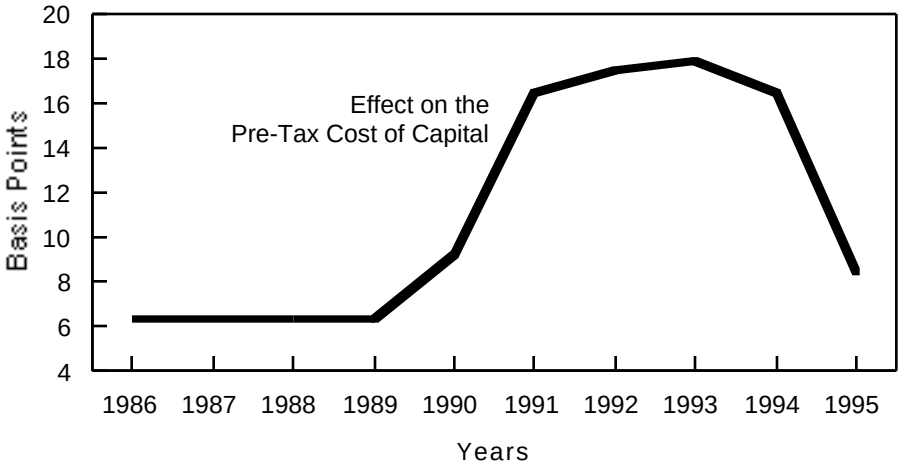
Figure 5 shows the marginal contribution of the cost of deposit insurance premiums to the pre-tax user cost of capital, calculated by subtracting a measure of the pre-tax user cost of capital that excludes deposit insurance premiums from the pre-tax user cost including deposit insurance premiums. The dramatic increase in the cost of deposit insurance after 1989 reflects the increases in effective deposit insurance assessments imposed during this period. These increases had a substantial impact on banks cost of capital. From 1992 to 1994, deposit insurance premiums contributed between 16 and 18 basis points to the pre-tax user cost of capital, up from 6 basis points in 1988.

Whether deposit insurance assessments should be treated as a tax on the banking industry depends on how fairly the FDIC's assessments reflect the cost of its financial guarantee. A recent study by Epps, Pulley, and Humphrey (1996) computed "fair" deposit insurance premiums for a sample of 77 banks using 1989 data. That study found that the median fair deposit insurance premium was 0.0107 percent of deposits (assuming one bank examination per year), compared to the 0.0833 deposit insurance premium charged that year. At first glance, these findings seem to suggest that deposit insurance is overpriced. A closer look at the authors' results reveals certain important mitigating factors, however. The fair deposit insurance premiums for individual banks in the study ranged from a low under 0.0001 percent to a high of 0.7749 percent. The authors note, however, that the FDIC can reduce the effective cost of its liability to depositors of a troubled bank through more frequent monitoring, which is the current policy of the bank regulatory agencies.[22] Nonetheless, these findings indicate that deposit insurance requires well-managed and conservatively run banks to subsidize banks that pose greater risks to the deposit insurance fund. To be sure, the adoption of risk-based assessments has ameliorated this problem somewhat. But although the adoption of risk-based assessments has reduced the subsidy to risky banks, the findings of Epps, Pulley, and Humphrey (1996) indicate that the current risk-based assessment scheme would not have been sufficient to eliminate the subsidy to the riskiest banks in 1989. What Figure 5 shows, then, is the deposit insurance tax on the safest banks. Based on available

[21] For more details, see FDIC (1995).

[22] The fair deposit insurance premiums reported here do not include the cost of bank examinations. Epps, Pulley, and Humphrey (1996) also discuss examination costs.

Figure 5 The Cost of Deposit Insurance



information, it is not clear whether the current risk-based assessment scheme constitutes a tax on the industry as a whole.

Legislation enacted during 1996 does use the deposit insurance system to impose a tax on commercial banks, however. Beginning in 1997, all banks will be required to pay a special charge of 1.29 basis points to help pay the interest on bonds issued in 1987 to recapitalize the thrift industry's deposit insurance fund. That surcharge is scheduled to increase to 2.43 basis points in 1999. Understandably, the commercial banking industry resisted legislation requiring it to help pay for the losses incurred by the thrift industry. The banking industry had paid very high deposit insurance premiums during the early 1990s to recapitalize its own deposit insurance fund, and bankers did not wish to see their premiums raised once again to help rescue a competing industry's fund. My model shows that these surcharges will not have a dramatic impact on the banking industry's cost of capital, however. Using 1995 year-end data, the effect of a 1.29 basis point surcharge would be to increase the pre-tax user cost of capital by less than 1 basis point. A 2.43 basis point surcharge would produce an increase of less than 2 basis points.

4. CONCLUDING COMMENTS

In 1986, the banking industry paid an average effective tax rate of just under 24 percent. Since the enactment of the Tax Reform Act of 1986, however, the average effective tax rate has been over 30 percent. A cursory inspection of these data would seem to suggest that the tax burden on the banking industry

has risen in recent years. A closer look at the factors accounting for this rise suggest otherwise, however. The increase in the average tax rate paid by banks over the past decade is due largely to the elimination of the interest deduction on municipal debt for banks. Banks, in return, have responded by substituting into taxable corporate debt, which pays higher interest rates. Although banks now pay more federal taxes, they also earn more pre-tax income. Moreover, the elimination of the tax deduction for municipal debt had no impact on banks' incentive to extend other forms of credit, which is influenced by the marginal effective tax rate on bank lending.

An examination of the recent behavior of the marginal effective tax rate on bank lending paints a much different picture, suggesting that the tax disincentives to commercial bank intermediation have fallen modestly over the past ten years. The decline in the marginal effective tax rate is due principally to two factors. The first is the Tax Reform Act of 1986. Although tax reform resulted in higher average tax rates, it reduced the marginal effective tax rate on commercial bank lending. The second factor is the reduction in the implicit reserve requirement tax, which is due partly to reductions in reserve requirements and partly to declining interest rates.

This article began by questioning the extent to which the tax burden borne by the commercial banking industry may have contributed to the declining share of bank lending in credit markets. For many years, the commercial banking industry enjoyed special tax treatment, meant in part to compensate for the burden of regulation, including the cost of reserve requirements. Commercial banks now face the same federal tax rules as other lenders, however. At the same time, they also continue to bear the cost of reserve requirements. Although reserve requirements have been reduced in recent years, they still impose an approximately 10 basis point cost penalty on banks out of a total marginal effective tax rate of roughly 70 basis points. Even though the tax burden on commercial banking has fallen by some measures, implicit taxes continue to handicap the ability of banks to compete against other lenders. More importantly, recent statutory reductions in reserve requirements accounted for less than half of the reduction in the cost of reserve requirements in recent years—the rest was due to falling interest rates. In the absence of further policy actions, then, an increase in interest rates could increase the marginal effective tax rate on commercial bank lending substantially.

Many observers feel that any regulatory burden borne by banks, including the burden of reserve requirements, is mitigated by unique benefits such as deposit insurance and access to the Fed's discount window. The foregoing analysis showed that changes in deposit insurance assessments contributed substantially to the banking industry's cost of capital from 1992 to 1995. Deposit insurance assessments have fallen dramatically over the past year, however, and now account for a negligible fraction of the cost of financing a loan (except for the few banks that must pay the highest deposit insurance assessment rate of 31

basis points). Moreover, the estimated impact of the recent deposit insurance surcharge imposed on commercial banks to help pay for the recapitalization of the thrift industry's deposit insurance fund is exceedingly small and would appear to pose no undue burden on the industry. Whether deposit insurance represents a subsidy to the banking industry continues to be the topic of an active debate. Fortunately, the burden of explicit corporate taxes and the implicit cost of reserve requirements can be quantified.

APPENDIX : ESTIMATION OF THE USER COST OF CAPITAL

For the financial cost of capital, $\rho = \sum_{i=1}^{5} \lambda_i \rho_i$, estimates of interest expense were obtained using data available in the *FDIC Historical Statistics on Banking*. The cost of equity capital was estimated using the basic CAPM model, following the procedure suggested by Ibbotson and Sinquefield (1989). The estimate for the stock market beta needed for this calculation was obtained from Berkovec and Liang (1991). The results are summarized below.

THE AVERAGE COST OF EQUITY CAPITAL U.S. COMMERCIAL BANKS

Year	Risk-Free Interest Rate	Beta	Average Equity Premium	Cost of Equity (ρ_5)
1986	6.16%	0.95	0.88	14.52%
1987	5.47%	0.95	0.88	13.83%
1988	6.35%	0.95	0.88	14.71%
1989	8.37%	0.95	0.88	16.73%
1990	7.81%	0.95	0.88	16.17%
1991	5.60%	0.95	0.88	13.96%
1992	3.51%	0.95	0.88	11.87%
1993	2.90%	0.95	0.88	11.26%
1994	3.90%	0.95	0.88	12.26%
1995	5.60%	0.95	0.88	13.96%

Estimates of financial structure, as reflected by the parameters $\lambda_1, \lambda_2, \ldots, \lambda_5$, were obtained from the Quarterly Reports of Condition and Income, or Call Reports, and from the Federal Reserve's Weekly Report of Transaction Accounts (FR2900). Data on loan charge-off rates came from the *FDIC Historical Statistics on Banking*, while data on effective deposit insurance assessments are from the FDIC *Annual Report* for 1995.

REFERENCES

Berkovec, James A., and J. Nellie Liang. "Changes in the Cost of Equity Capital for Bank Holding Companies and the Effects on Raising Capital." Finance and Economics Discussion Series 160. Washington: Board of Governors of the Federal Reserve System, 1991.

Bradford, David F., and Don Fullerton. "Pitfalls in the Construction and Use of Effective Tax Rates," in Charles R. Hulten, ed., *Depreciation, Inflation, and the Taxation of Income from Capital.* Washington: The Urban Institute Press, 1981.

Buynak, Thomas M. "How Will Tax Reform Affect Commercial Banks?" Federal Reserve Bank of Cleveland *Economic Review* (Quarter 2, 1987), pp. 24–34.

Epps, T. W., Lawrence B. Pulley, and David B. Humphrey. "Assessing the FDIC's Premium and Examination Policies Using 'Soviet' Put Options," *Journal of Banking and Finance*, vol. 20 (May 1996), pp. 699–721.

Federal Deposit Insurance Corporation. *Annual Report 1995.* Washington: 1995.

Fullerton, Don. "Which Effective Tax Rate?" *National Tax Journal,* vol. 37 (March 1984), pp. 23–41.

_____, and Yolanda K. Henderson. "Long-Run Effects of the Accelerated Cost Recovery System," *Review of Economics and Statistics,* vol. 67 (August 1985), pp. 363–72.

Hall, Robert E., and Dale W. Jorgenson. "Tax Policy and Investment Behavior," *American Economic Review,* vol. 57 (June 1967), pp. 391–414.

Harberger, Arnold C. "Efficiency Effects of Taxes on Income from Capital," in M. Krzyzaniak, ed., *Effects of Corporation Income Tax.* Detroit: Wayne State University Press, 1966.

Henderson, Yolanda. "The Taxation of Banks: Particular Privileges or Objectionable Burdens?" *New England Economic Review* (May/June 1987), pp. 3–18.

Ibbotson, Roger G., and Rex A. Sinquefield. *Stocks, Bonds, Bills, and Inflation: Historical Returns (1926–1987).* Charlottesville, Va.: The Research Foundation of The Institute of Chartered Financial Analysts, 1989.

McCauley, Robert N., and Rama Seth. "Foreign Bank Credit to U.S. Corporations: The Implications of Offshore Loans," Federal Reserve Bank of New York *Quarterly Review,* vol. 17 (Spring 1992), pp. 52–65.

Neubig, Thomas S. "The Taxation of Financial Institutions After Deregulation," *National Tax Journal,* vol. 37 (September 1984), pp. 351–59.

_____, and Martin A. Sullivan. "The Effect of the Tax Reform Act of 1986 on Commercial Banks," in *Compendium of Tax Research 1987.* Washington: Office of Tax Analysis, Department of the Treasury, 1987.

O'Brien, James M., and Matthew D. Gelfand. "Corrigendum: The Impact of the Tax Reform Act of 1986 on Commercial Banks," *Tax Notes*, vol. 34 (March 30, 1987a), pp. 1323–25.

_____. "Effects of the Tax Reform Act of 1986 on Commercial Banks," *Tax Notes,* vol. 34 (February 9, 1987b), pp. 597–604.

U.S. Congress, Joint Committee on Taxation. *General Explanation of the Tax Reform Act of 1986*. Public Law 99–514, 100 Cong. 1 Sess. Washington: Government Printing Office, 1987.

Walter, John R. "Loan Loss Reserves," Federal Reserve Bank of Richmond *Economic Review,* vol. 77 (July/August 1991), pp. 20–30.